The amendments to claims 38, 39, and 41 are supported by the present specification. For example, exemplary support for the term "fraction" is found in Example 12 and is shown in Figure 12. Additionally, support for "cell free medium" is found on page 34, line 10, and page 38, line 3.

Because the foregoing amendments do not introduce new matter, entry thereof by the Examiner is respectfully requested.

II. Summary of the Invention

The present invention is directed to cell free mediums comprising a fusion protein of hPTH (1-84) and a leader sequence, and a method for obtaining intact hPTH (1-84) using such a cell free medium.

PTH, which is an important regulator of calcium metabolism in mammals, is related to several mammalian diseases, such as milk fever, acute hypocalsemia, and otherwise pathologically altered blood calcium levels. *See* page 2, lines 15-19, of the application. Through its action on target cells in bone and kidney tubuli, PTH increases serum calcium and decreases serum phosphate, while opposite effects are found regarding urinary excretion of calcium and phosphate. *See* page 5, lines 27-31, of the application. PTH is useful, for example, as a component of a diagnostic kit or as a therapeutic in human and veterinary medicine. *See* page 2, lines 20-22, of the specification.

Prior to the present invention, PTH was commercially available only in very small quantities at high cost, partly because synthesis of the compound was difficult and complex. *See* page 1, lines 33-38, of the specification. Applicants have overcome the problems of the prior art and discovered processes of preparing recombinant PTH using microorganisms.

III. Issues Under Objections

The Examiner objected to the specification for allegedly failing to provide antecedent basis in the specification for the recitation of "greater than 90%" purity of the claimed

protein. Applicants note that the present specification provides antecedent basis for this recitation and direct the Examiner's attention to page 7, lines 30-32, where it states: "The sequence analysis indicated that the recombinant hPTH was more than 90 percent pure." Therefore, Applicants respectfully request reconsideration and withdrawal of the objection.

The Examiner also objected to the specification because the information on page 1, lines 20-23, appears to be duplicative of the first paragraph of the same page. Applicants have amended the specification by deleting lines 20-23 on page 1.

IV. Claim Rejections - 35 U.S.C. § 112, First Paragraph

Claims 34 and 52 are rejected by the examiner under 35 U.S.C. § 112, first paragraph, for alleged lack of enablement. Applicants have canceled claims 34 and 52, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

Claims 24-29, 34-47, and 52-56 were rejected by the examiner under 35 U.S.C. § 112, first paragraph, for alleged lack of enablement. The examiner asserted that the specification provides enablement for Saccharomyces mating factor α1 leader sequence for use in yeast; however, the specification allegedly does not provide enablement for all leader sequences for use in all microorganisms. Applicants respectfully request reconsideration and withdrawal of the rejection.

Applicants have canceled claims 24-29, 34-37, 40, 42-47, and 52-56, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

A. Numerous Signal Sequences were Characterized and Well Known in the art at the Time the Claimed Invention was Made

With respect to claims 38, 39, and 41, the attached articles demonstrate that the use of various leader sequences in various microorganisms was well known in the art at the time the claimed invention was made. For example, Von Heune, "Patterns of Amino Acids Near

Signal-Sequence Cleavage Sites," *Eur. J. Biochem., 133*:17-21 (1983) (Exhibit 1), teaches that the patterning of amino acids near the cleavage site between a signal sequence and a mature protein "is much richer than hitherto believed, and that, given this information, cleavage sites can be predicted quite successfully." Abstract of Von Heune. This reference presents an analysis of a comprehensive collection of known eukaryotic signal sequences, aligned with coincident cleavage sites (Table 1).

In addition, von Heijne, "Signal Sequences; The Limits of Variation," *J. Mol. Biol., 184*:99-105 (1985) (Exhibit 2), teach mapping of a large sample of N-terminal and C-terminal regions of signal sequences, including eukaryotic and prokaryotic signal sequences. This reference further teaches that "[a]ll signal sequences seem to be built along the same general lines . . ." Col. 1, page 99. The analysis of eukaryotic and prokaryotic signal sequences shows subtle differences between the two types, and suggests "minimal" requirements to which a fully functional signal sequence must conform. *Id.*

Moreover, Watson, "Compilation of Published Signal Sequences," *12*(13):5145-5164(1984) (Exhibit 3), teaches that the structure, function, and processing of signal sequences has been reviewed (page 5146), and follows with a compilation of published signal sequences (pp. 5146-5157). The proteins listed are grouped into genera, the proteins are classified as inner membrane, outer membrane, periplasmic, or transmembrane, where appropriate, and the first ten amino acid residues of the protein sequences are also given.

About 189 eukaryotic signal sequences are listed, including sequences from baboon, bovine, canine, hamster, human, monkey, murine, ovine, porcine, rabbit, rat, chicken, caiman crocodylus xenopus, angler fish carp, catfsh, magfish, salmon, torpedo, winter flounder, bee, *Drosophila melanogaster*, *Brasica napus*, *Hordeum vulgare*, *Phaseolus vulgaris*, *Pisum sativum, Zea maize,* Saccharomyces, *Plasmodium knowlesi*, and *Trypandsoma brucei*. Over 24 viral signal sequences are listed, including sequences from adeno virus 2, herpes simplex, avian influenza, human influenza, mouse mammary tumour virus, rabies virus, roue sarcoma virus, simian rotavirus-SA11, vesicular stomatitis virus – hamster, vesicular stomatitis virus – human, and yeast killer. Over 55 prokaryotic signal sequences are listed, including sequences

from *Bacillus amyloliquitagiens, Bacillus cereus, Bacillus licheniformis, Bacillus subtilis, Bacteroides modosus, Corynebacterium ditheriae, Escherichia coli, Zrwinia Amylovora, Enterobacter aerogenes, Halobacterium Malonium, H. morganii, Moraxella nonliquefaciens, Neisseria gonorrhoeae, Pseudomonas sp., Salmonella typhimurium, Serratia hargescens, Shigella dysenteriae Staphlococcus aureus and Vibrio cholerae.*

Thus, the characterization of numerous signal sequences was well known in the art at the time the claimed invention was made.

B. Numerous Expression Systems Were Known in the art at the Time the Claimed Invention was Made

The cited references (Exhibits 1, 2, and 3) also identify a wide variety of expression systems that can be used for making recombinant proteins, including eukaryotic, prokaryotic, and viral expression systems.

As signal sequences other than Saccharomyces mating factor alpha-1, and expression systems other than yeast, were well known in the art at the time the claimed invention was made, Applicants' claimed invention is not limited to these specific embodiments to satisfy the enablement requirement of 35 U.S.C. § 112, first paragraph. Withdrawal of this ground for rejection is respectfully requested.

V. Claim Rejections - 35 U.S.C. § 112, Second Paragraph

Claim 42 was rejected by the Examiner under 35 U.S.C. § 112, second paragraph, for reciting "STE13 recognition". Applicants have canceled claim 42, thus obviating the rejection of this claim. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

Claims 24-29, 34-37, 38-47, and 52-56 were rejected by the Examiner under 35 U.S.C. § 112, second paragraph, because it is allegedly unclear how and where the leader sequence is cleaved in microorganisms other than yeast. Applicants have canceled claims 24-29, 34-47, and 52-56, thus obviating the rejection of these claims. The cancellation of

claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

With respect to claims 38, 39 and 41, Applicants have detailed above in Section IV the knowledge in the art at the time the claimed invention was made regarding signal sequences and expression systems. In particular, it is noted that von Heune (Exhibit 1) teaches that using the cleave prediction method described in the reference, only 5 out of 76 processing sites were incorrectly predicted. Col. 2, page 17. Thus, use of various signal sequences in various expression systems does not render the invention unpredictable or indefinite. For at least these reasons, withdrawal of the ground for rejection is respectfully requested.

Claims 38, 41, 42, 45, 48, 52, and 54 were rejected by the Examiner under 35 U.S.C. § 112, second paragraph, because the examiner asserts that the term "component" is confusing. Applicants have canceled claims 42, 48, 52 and 54, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

With respect to claims 38 and 41, Applicants have replaced the term "component" with the term "fraction". The fraction is a result of the column chromatography that is described in Example 12. Therefore, support for "fraction" is found in the specification in Example 12 and is shown in Figure 12. Additionally, the term "fraction" is a well-recognized term of art.

Claims 44 and 47 were rejected by the Examiner under 35 U.S.C. § 112, second paragraph, because claims 44 and 47 should allegedly refer to part (d) instead of part (c) of claims 43 and 45, respectively. Applicants have canceled claims 44 and 47, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

Claims 24-29, 34-37, 42-47, and 52-56 were rejected by the Examiner under 35 U.S.C. § 112, second paragraph, because it is allegedly not clear if the PTH is purified before

or after it becomes part of the claimed composition. Applicants have canceled claims 24-29, 34-37, 42-47, and 52-56, thus rendering the rejection of these claims moot. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

VI. Claim Rejections - 35 U.S.C. § 102

Claims 24-29, 34-37, 42-47, and 52-56 were rejected by the Examiner under 35 U.S.C. § 102 as being unpatentable over either Keutmann et al. or Kimura et al. Applicants have canceled claims 24-29, 34-37, 42-47, and 52-56, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

Claims 24-29, 34-37, 42-47, and 52-56 were rejected by the examiner under 35 U.S.C. § 102(b) as being allegedly anticipated by either Brewer et al. (U.S. Patent No. 3,886,132), Kumagaye et al. (*J. Chrom.*, 327:327), or Fairwell et al. (*Biochem,* 22:2691). Applicants have canceled claims 24-29, 34-37, 42-47, and 52-56, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

VII. Rejections Under 35 U.S.C. § 102/103

Claims 24-29, 34-37, 42-27, and 52-56 were rejected by the Examiner under 35 U.S.C. § 102 as being allegedly anticipated by, or under 35 U.S.C. § 103 as being allegedly obvious over, Brewer et al. '132. Applicants have canceled claims 24-29, 34-37, 42-47, and 52-56, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

Claims 24-29, 34-37, 42-47, and 52-56 were rejected as being allegedly anticipated under 35 U.S.C. § 102 by, or allegedly obvious under 35 U.S.C. § 103 over, Kumagaye et al., Kimura et al., or Fairwell et al. Applicants have canceled claims 24-29, 34-37, 42-47, and

52-56, thus obviating the rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

VIII. Provisional Obviousness-type Double Patenting Rejection Under 35 U.S.C. § 101

Claims 24-29, 34-37, 42-47, and 52-56 were provisionally rejected under the judicially created doctrine of obviousness-type double patenting over claims 31-35 of copending Application No. 08/340,664. Applicants have canceled claims 24-29, 34-37, 42-47, and 52-56, thus obviating the provisional rejection of these claims. The cancellation of claims does not constitute acquiescence in the propriety of any rejection set forth by the Examiner.

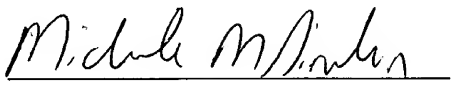
CONCLUSION

As the above-presented amendments and remarks address and overcome all of the rejections presented by the Examiner, withdrawal of the rejections and allowance of the claims are respectfully requested.

If the Examiner has any questions concerning this application, he or she is requested to contact the undersigned.

Respectfully submitted,

Date July 15, 2002

By 

FOLEY & LARDNER
Washington Harbour
3000 K Street, N.W., Suite 500
Washington, D.C. 20007-5109
Telephone: (202) 672-5538
Facsimile: (202) 672-5399

Michele M. Simkin
Attorney for Applicant
Registration No. 34,717

Should additional fees be necessary in connection with the filing of this paper, or if a petition for extension of time is required for timely acceptance of same, the Commissioner is hereby authorized to charge Deposit Account No. 19-0741 for any such fees; and applicant(s) hereby petition for any needed extension of time.

VERSION WITH MARKINGS TO SHOW CHANGES MADE

38. (Amended) A cell free medium [An extract] obtained following growth of a microorganism transformed to express a DNA sequence encoding a fusion product in which hPTH(1-84) is fused at its N-terminus with a leader sequence, wherein:

(a) said leader sequence is cleavable by said microorganism upon production of said fusion product;

(b) said cell free medium [extract] comprises a PTH fraction [component]; and

(c) said PTH fraction [component] consists essentially of intact [PTH] hPTH(1-84) molecules.

39. (Amended) The cell free medium [An extract] according to claim 38, wherein the PTH fraction [component] consists of [PTH molecules that are] intact [PTH(1-84)] hPTH(1-84) molecules.

41. A method for obtaining intact hPTH(1-84), comprising

(a) obtaining [an extract following growth of a microorganism transformed to express a DNA sequence encoding a fusion product in which hPTH(1-84) is fused at its N-terminus with a leader sequence, wherein] the cell free medium of claim 38[:

(i) said leader sequence is cleavable by said microorganism upon production of said fusion product;

(ii) said extract comprises a PTH component; and

(iii) said PTH component consists essentially of intact PTH molecules]; and

(b) treating said [extract] medium to isolate [the PTH component thereof] intact hPTH(1-84).

Eur. J. Biochem. 133, 17–21 (1983)
© FEBS 1983

Patterns of Amino Acids near Signal-Sequence Cleavage Sites

Gunnar VON HEIJNE
Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, Stockholm

(Received October 27/December 30, 1982) – EJB 6142

According to the signal hypothesis, a signal sequence, once having initiated export of a growing protein chain across the rough endoplasmic reticulum, is cleaved from the mature protein at a specific site. It has long been known that some part of the cleavage specificity resides in the last residue of the signal sequence, which invariably is one with a small, uncharged side-chain, but no further specific patterns of amino acids near the point of cleavage have been discovered so far. In this paper, some such patterns, based on a sample of 78 eukaryotic signal sequences, are presented and discussed, and a first attempt at formulating rules for the prediction of cleavage sites is made.

The molecular basis for the highly precise cleavage between signal sequences and mature exported proteins has been somewhat of an enigma ever since the first pre-proteins were sequenced some ten years ago [1]. The last residue of the signal sequence has been found to be either Ala, Gly, Ser, Cys, Thr or Gln, i.e. one with a small, uncharged side-chain, and it has also been shown that the signal sequence has a very hydrophobic central core [2], but neither the positioning of this core relative to the cleavage site, nor the patterns of amino acids found at this site have so far allowed a precise prediction of the position of the point of cleavage from sequence data alone. In this paper, we show that the patterning of amino acids near the cleavage site is much richer than hitherto believed, and that, given this information, cleavage sites can be predicted quite successfully.

RESULTS AND DISCUSSION

Analysis of a comprehensive collection of known eukaryotic signal sequences, aligned with coincident cleavage sites (Table 1), reveals some interesting regularities. As shown in Table 2, whole classes of residues are absent not only from position −1 (the last residue of the signal sequence, cf. above) but also from position −3, and, conversely, the same types of amino acids as in position −1 are strongly preferred in position −3. Also, these small, neutral residues are quite rare in position −2 (there is only one Ala in position −2, in contrast to position −1 with 36, and position −3 with 25 alanines).

Moreover, the bulky aromatic residues seem to be selected for in position −2, whereas the secondary-structure-disrupting residues Gly and Pro are found predominantly in positions −4 and −5, respectively. In addition, there is not a single Pro residue in the whole region between −3 and +1.

The hydrophobic residues seem to be merely tolerated in positions +1 and −2 to −4; they are conspicuously rare in position −5, and then become dominant from position −6 onwards as one goes further into the central hydrophobic core (see above).

Finally, although charged residues are tolerated to some extent in positions −2, −4 and −5, they are much more abundant in the mature sequences, starting from position +1.

A similar analysis of a smaller sample with all isozymes and obviously related sequences removed yields essentially identical results (data not shown).

These observations make it possible to formulate a preliminary set of rules that can be used to predict the point of cleavage in a given pre-protein sequence. At this stage, we have felt it reasonable to deal with classes of amino acids rather than individual ones, and the following scheme can thus be refined as more sequences and data for individual residues become available.

The prediction method proposed here has two components: first, the N terminus of the hydrophobic core is located by searching from the N terminus for the first quadruplet with at least three hydrophobic residues (group IV in Table 2, not counting the initiator Met), and defining a 'window' for processing between residues $i + 12$ and $i + 20$ (i being the first residue in the quadruplet). Then, for each residue in the 'window', a measure of 'processing probability' is calculated by multiplying together the respective values from Table 3, values chosen 'by inspection' so as roughly to reflect the amino acid patterns discussed above, and fine-tuned to maximize the number of correct predictions. Finally, the site of processing is predicted as the site with the highest 'processing probability'.

Applying this scheme to the sequences listed in Table 1, only 5 out of 76 processing sites are incorrectly predicted (nos 18, 22, 37, 42, 45), with one additional case where the method cannot differentiate between the correct and one incorrect site (no. 52).

Using less than the full set of rules, it was found that the rule 'no Pro in position +1' can be discarded without affecting the number of correct and undecided predictions for the present data base. All other rules are important, though. As an illustration, using only the rule for position −1 yields 16 incorrect and 39 undecided predictions, and adding the rule for position −3 only improves these figures to 14 incorrect and 16 undecided predictions.

The amino acid patterns presented here can perhaps also yield some insight into the actual structure of the signal-sequence–protease complex. As shown above, positions −1 and −3 seem to be strongly selected for small, neutral residues, whereas positions +1, −2, and −4 seem to accomodate almost any kind of residue. Moreover, there is not a single Pro from position −3 to +1, and, further, this region is often separated from the hydrophobic core of the signal sequence by the strong helix-breakers Pro or Gly [3].

Thus, a possible signal-sequence–protease structure would be an alternating sheet structure near the cleavage site, joined at its N terminus to a largely hydrophobic helix (Fig. 1). Possibly, position +1 is located close to the membrane surface, as

Table 1. *A collection of eukaryotic signal sequences*

In the first part, the sequences have been aligned from their known cleavage sites (between positions −1 and +1); in the second part sequences in which the location of this site is unknown have been aligned from their predicted sites of cleavage (see text). The predicted cleavage sites are indicated by a star (*), and the processing 'window' is shown by slashes (/; when a * and a / coincides, only the * is shown). Amino acids are symbolized by the one-letter code, i.e. A = Ala, C = Cys, D = Asp, E = Glu, F = Phe, G = Gly, H = His, I = Ile, K = Lys, L = Leu, M = Met, N = Asn, P = Pro, Q = Gln, R = Arg, S = Ser, T = Thr, V = Val, W = Trp, Y = Tyr, X = unknown. Proteins are as follows; 1: whey phosphoprotein, rat; 2: α-1 acid glycoprotein, rat; 3: α-thyrotropin, mouse; 4: insulin, hagfish; 5: insulin, anglerfish; 6: insulin, human; 7: insulin I, rat; 8: insulin II, rat; 9: β-casein, ovine; 10: κ-casein, ovine; 11: α-lactalbumin, ovine; 12: β-lactoglobulin, ovine; 13: αs1 casein, ovine; 14: αs2 casein, ovine; 15: glycoprotein, VS virus; 16: VLDL-II, cockerel; 17: melittin, bee; 18: lactin, rat; 19: placental lactogen, human; 20: β-choriogonadotropin, human; 21: α-choriogonadotropin, human; 22: uteroglobin, rabbit; 23: growth hormone, rat; 24: growth hormone, human; 25: growth hormone, bovine; 26: parathyroid hormone, bovine; 27: relaxin, rat; 28: serum albumin, rat; 29: serum albumin, human; 30: liver albumin, rat; 31: tropoelastin H, chicken; 32: ovomucoid, chicken; 33: lysozyme, chicken; 34: conalbumin, chicken; 35: α-1 antitrypsin, human; 36: prostatic binding protein, rat; 37: prostatic binding protein C3, rat; 38: glycoprotein, rat; 39: apolipoprotein A1, rat; 40: glycoprotein, rabies virus; 41: hemagglutinin, human influenza Victoria; 42: hemagglutinin, human influenza Jap; 43: hemagglutinin, avian influenza FPV; 44–50: leukocyte interferon, human; 51: immune interferon, human; 52: fibroblast interferon, human; 53–56: κ-immunoglobulin, mouse; 57–59: λ-immunoglobulin, mouse; 60, 61: κ-immunoglobulin, mouse; 62: H-chain immunoglobulin, mouse; 63–66: embryonic VH-immunoglobulin, mouse; 67: H-chain immunoglobulin, mouse; 68: trypsinogen 1, canine; 69: trypsinogen 2+3, canine; 70: chymotrypsinogen 2, canine; 71: carboxypeptidase A1, canine; 72: amylase, canine; 73, 74: amylase, mouse; 75: amylase, rat; 76: α-lactalbumin, rabbit; 77: α-lactalbumin, porcine; 78: carboxypeptidase A, rat; 79: ACTH-β-LPH precursor, bovine; 80: ACTH-β-LPH precursor, porcine; 81: ACTH-β-LPH precursor, human; 82: gastrin, porcine; 83: renin, mouse; 84: glycoprotein, trypanosome; 85: somatostatin, catfish; 86–88: somatostatin, anglerfish; 89: calcitonin, rat; 90: glucagon, anglerfish

| Protein | Amino acid in position | | |

Table 1 (continued)

Protein	Amino acid in position −20	−15	−10	−5	1*1	5	10	Reference
47	M A S P F A L L M V L V V L S C/K S S C S L G*C/D L P E T H S L G							[40]
48	M A L S F S L L M A V L V L S Y/K S I C S L G*C/D L P Q T H S L G							[40]
49	M A L P F S L M M A L V V L S C/K S S C S L G*C/N L S Q T H S L N							[40]
50	M A L P F A L M M A L V V L S C/K S S C S L G*C/N L S Q T H S L N							[41]
51	M K Y T S Y I L A F Q L C I V L G S/L G*C Y C Q D P/Y V K E							[42]
52	[illegible]							[43]
53	M R A P A Q I F G F L L L L F P G T/R C*D I Q M T Q/S P S S							[44]
54	M E T D T L L L W V L L L W V P/G S T G*D I V L/T Q S P A S							[44]
55	M E T D T L L L W V L L L W V P/G S T G*N I V L/T Q S P A S							[44]
56	M E T D T L L L W V L L L W V P/G A D A*A P T V/S I F P P S							[44]
57	M A W I S L I L S L L A L S S/G A I S*Q A V V/T Q E S A L							[44]
58	M A W I S L I L S L L A L S S/G A I S*Q A V V/T Q E S A L							[44]
59	M A W T S L I L S L L A L C S G/A S S*Q A V V T/Q E S A L							[44]
60	M G V R M E S H T R V P I F L L L W L S G/T D G*D I V X X/X X X X X							[44]
61	M D M R A P A Q I F G F L L L L F P G T/R C*D I Q M T Q/S P S S							[44]
62	M K V L S L L Y L L T A I P/G I M S*D V Q L/Q E S G P G							[44]
63	M G W S W I F L F L L L S G T A G/V H S*X X X X X/X X X X X							[45]
64	I K W S W I S L F L L L S G T A G V/H S*X X X X X X/X X X X							[45]
65	M E C S W V F L F L L L S L T A G/I H C*X X X X X/X X X X X							[45]
66	M E W S G V F I F L L L S V T A G/V Y S*X X X X X/X X X X X							[45]
67	M G W S F I P L F L L L S V T A/G V H S*E V Q L/Q Q S G A E							[46]
68	X X P L L I L A F L X A A V/A*T P T D D D D/K I V							[47]
69	A L X I I F L A L L X A X/V A*F P I D D D/D K I V							[47]
70	A F L I L V X A F A L X X V A F X X X V P A I X P V X							[47]
71	X X L I L V F G A L L X A I Y X Q X A F V X X X X X							[47]
72	X X F F L L L X V I G F X/V A*Q Y A P H X/X X X X							[47]
73	M K F F L L L S L I G F C/W A*Q Y D P H T/Q Y G R							[47]
74	M K F V L L L S L I G F C/W A*Q Y D P H T/S D G R							[47]
75	L L S L I G F C Y A*Q Y D P H T A D G R							[47]
76	M M P L V P L L L V S I V/F P G I Q A*T Q/L T R C E L T E							[48]
77	M M S F V S L L V V G I L/F P A I Q A*K Q/F T K C E L S Q							[48]
78	M K R L L I L S L L L E A V/C G*N E N F V G/H Q V L							[49]
	Sequences with unknown cleavage sites							
79	M P R L C S S R S G A L L L A L L L Q A S M/E V R G*W C L E/S S Q C Q D							[50]
80	M A W Q G L L L A A C L L V L P/S T M A*D C L S/G C S L C A							[51]
81	M A R P L T L C T W L L L L G/P G L L A*T V R/A R C S Q D C							[52]
82	M Q R L C A Y V L I H V L A L A A C/S E A*S W K P G/F Q L X X							[53]
83	M D R R R M F L W A L L L L W S P C T F S*L P T G T T F E/R I							[54]
84	M V K A I A S L M L L H I W A I E E/I K A*E R Q A P/S V S R T							[55]
85	M S S S P L R L A L A L M C L V S A V G V/I S*C G R P H Y/V L N S							[56]
86	M V S S S R L R C L L V L L L S L T V S/I S C S F A*G Q/R D S K L R L L							[57]
87	M K M V S S S R L R C L L V L L L S L T A S/I S C S F A*G Q/R D S K L R L L							[58]
88	M Q C I R C P A I L A L L A L V L C G F/S V S*S Q L D R/E Q S D N							[58]
89	M G F L K F S F F L V V S I/L L L Y Q A C G*L Q A V P L R S T L							[59]
90	M K R I H S L A G I L L V L G L I Q S S/C*R V L M Q E A/D P S							[60]

indicated by the observed greater proportion of charged residues in this position (Table 2).

A structure of this kind would need some 8 or 9 helical residues added to the 4 or 5 in the extended sheet conformation to reach through the non-polar interior of the membrane, a figure that compares quite well with the length of the uncharged segment of the shortest known signal sequence, 13 residues.

A similar proposal has been made earlier on the basis of an analysis of predicted secondary structures for signal sequences [6] and the model is also in agreement with recent results from an energy-minimization study of the conformation of one particular signal sequence [5].

A signal sequence would thus contain two different and largely independent 'signals': one in the form of a hydrophobic core, presumably responsible for initiating export and for binding to the 'signal recognition protein' [6], and a second one in the region −5 to −1 conferring processing specificity.

This analysis has been based on a collection of eukaryotic signal sequences and it might be asked how well prokaryotic ones conform to the patterns presented here. Unfortunately, only some 16 prokaryotic signal sequences are known so that too great a reliance on a statistical analysis would hardly be justified at this point. From the limited data available it seems, however, that the only obvious differences are limited to a more pronounced exclusion of charged residues from the prokaryotic sequences (except, of course, at the N terminus), and a strong preference for small, neutral residues in position −6 (data not shown).

Table 2. *Number of residues of different physico-chemical character in positions −6 to +2 (cf. Fig. 1 and Table 1)*
The cleavage site is located between positions −1 and +1. I (aromatic residues): Phe, His, Trp, Tyr; II (charged residues): Asp, Glu, Lys, Arg; III (large polar residues): Asn, Gln; IV (hydrophobic residues): Phe, Ile, Leu, Met, Val; V (small neutral residues): Ala, Cys, Ser, Thr; G: Gly; P: Pro; *N.B.* Phe is included in groups I and IV since it is both bulky and aromatic (thus excluded from positions −1 and −3) and hydrophobic (thus abundant in the hydrophobic core)

Group	Number of residues in position							
	−6	−5	−4	−3	−2	−1	+1	+2
I	11	6	7	0	21	0	6	7
II	0	6	7	0	9	0	25	13
III	0	6	1	0	13	1	14	11
IV	42	7	19	22	29	0	13	16
V	27	27	22	52	9	57	18	9
G	0	7	21	1	3	18	2	4
P	7	16	6	0	0	0	0	11

Table 3. *'Statistical weights' of the various physico-chemical groups of amino acids in positions −5 to +1, used to calculate the 'processing probability' (see text)*
The groups are defined in the legend of Table 2

Position	'Statistical weight' of group						
	I	II	III	IV	V	G	P
+1	1.0	1.0	1.0	1.0	1.0	1.0	0.0
−1	0.0	0.0	0.5	0.0	Ala = 8.0 Ser = 4.0 Cys = 1.0 Thr = 1.0	4.0	0.0
−2	2.0	1.0	1.0	1.0	0.6	1.0	0.0
−3	0.0	0.0	0.0	1.0	3.0	1.0	0.0
−4	1.0	1.0	1.0	1.0	1.0	4.0	1.0
−5	1.0	1.0	1.0	0.7	1.0	1.0	4.0

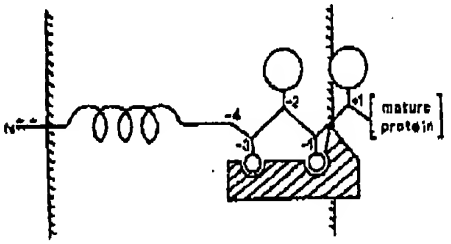


Fig. 1. *Proposed signal-sequence–protease complex.* The signal sequence spans the membrane as a 'helix + sheet' structure. The small, neutral residues in positions −1 and −3 fit into a pocket in the protease, thereby defining the cleavage site between positions −1 and +1

More generally, prokaryotic signal sequences are richer in Ala and poorer in Leu than eukaryotic ones [7].

In conclusion, it seems that a statistical approach to the problem of signal sequence processing can yield valuable information not readily obtainable from experimental work, and may, with suitable refinements, provide a good basis for schemes predicting the site at which processing of a given pre-protein sequence will take place.

This work was supported by a grant from the Swedish Natural Sciences Research Council.

REFERENCES

1. Strauss, A.W. & Boime, I. (1982) *CRC Crit. Rev. Biochem. 12*, 203–235.
2. von Heijne, G. (1982) *J. Mol. Biol. 159*, 537–541.
3. Chou, P. & Fasman, G.D. (1978) *Annu. Rev. Biochem. 47*, 251–276.
4. Austen, B.M. (1979) *FEBS Lett. 103*, 308–313.
5. Pincus, M.R. & Klausner, R.D. (1982) *Proc. Natl Acad. Sci. USA, 79*, 3413–3417.
6. Walter, P., Ibrahimi, I. & Blobel, G. (1981) *J. Cell Biol. 91*, 545–550.
7. von Heijne, G. (1981) *Eur. J. Biochem. 116*, 619–622.
8. Dandekar, A.M., Robinson, E.A., Appella, E. & Quasba, P.K. (1982) *Proc. Natl Acad. Sci. USA, 79*, 3987–3991.
9. Ricca, G.A. & Taylor, J.M. (1981) *J. Biol. Chem. 256*, 11199–11202.
10. Chin, W.W., Kronenberg, H.M., Dee, P.C., Maloof, F. & Habener, J.F. (1981) *Proc. Natl Acad. Sci. USA, 78*, 5329–5333.
11. Chan, S.J., Emdin, S., Kwok, S.C.M., Kramer, J.M., Falkmer, S. & Steiner, D.F. (1981) *J. Biol. Chem. 256*, 7595–7602.
12. Hobart, P.M., Shen, L.-P., Crawford, R., Pictet, R.L. & Rutter, W.J. (1980) *Science (Wash. DC) 210*, 1360–1363.
13. Bell, G.I., Swain, W.F., Pictet, R., Cordell, B., Goodman, H.M. & Rutter, W.J. (1979) *Nature (Lond.) 282*, 525–527.
14. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell, 18*, 545–558.
15. Mercier, J.-C. & Gaye, P. (1980) *Ann. N.Y. Acad. Sci. 343*, 232–251.
16. Lingappa, V.R., Katz, F.N., Lodish, H.F. & Blobel, G. (1978) *J. Biol. Chem. 253*, 8667–8670.
17. Chan, L., Bradley, W.A. & Means, A.R. (1980) *J. Biol. Chem. 255*, 10060–10063.
18. Molloy, C., Vlias, U. & Kreil, G. (1982) *Proc. Natl Acad. Sci. USA, 79*, 2260–2263.
19. McKean, D.J. & Maurer, R.A. (1978) *Biochemistry, 17*, 5215–5219.
20. Sherwood, L.M., Burstein, Y. & Schechter, I. (1979) *Proc. Natl Acad. Sci. USA, 76*, 3819–3823.
21. Birken, S., Fetherston, J., Canfield, R. & Boime, I. (1981) *J. Biol. Chem. 256*, 1816–1823.
22. Fiddes, J.C. & Goodman, H.M. (1979) *Nature (Lond.) 281*, 351–356.
23. Mulsky, M.L., Bullock, D.W., Willard, J.J. & Ward, D.N. (1979) *J. Biol. Chem. 254*, 1580–1585.
24. Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. & Goodman, H.M. (1977) *Nature (Lond.) 270*, 486–494.
25. Martial, J.A., Hallewell, R.A., Baxter, J.D. & Goodman, H.M. (1979) *Science (Wash. DC) 205*, 602–606.
26. Miller, W.L., Martial, J.A. & Baxter, J.D. (1980) *J. Biol. Chem. 255*, 7521–7524.
27. Kronenberg, H.M., McDevitt, B.E., Majzoub, J.A., Nathans, J., Sharp, P.A., Potts, P.A. & Rich, A. (1979) *Proc. Natl Acad. Sci. USA, 76*, 4981–4985.
28. Hudson, P., Haley, J., Cronk, M., Shine, J. & Niall, H. (1981) *Nature (Lond.) 291*, 127–131.
29. Sargent, T.D., Yang, M. & Bonner, J. (1981) *Proc. Natl Acad. Sci. USA, 78*, 243–246.
30. Dugaiczyk, A., Law, S.W. & Dennison, O.E. (1982) *Proc. Natl Acad. Sci. USA, 79*, 71–75.
31. Strauss, A.W., Bennett, C.D., Donohue, A.M., Rodkey, J.A. & Alberts, A.W. (1977) *J. Biol. Chem. 252*, 6846–6855.
32. Karr, S.R. & Foster, J.A. (1981) *J. Biol. Chem. 256*, 5946–5949.
33. Palmiter, R.D., Thibodeau, S.N., Rogers, G. & Boime, I. (1980) *Ann. N.Y. Acad. Sci. 343*, 192–209
34. Leicht, M., Long, G.L., Chandra, T., Kurachi, K., Kidd, V.J., Mace, M., Davie, E.W. & Woo, S.L. (1982) *Nature (Lond.) 297*, 655–659.
35. Parker, M., Needham, M. & White, R. (1982) *Nature (Lond.) 298*, 92–94.
36. Persson, H., Jörnvall, H. & Zabielski, J. (1980) *Proc. Natl Acad. Sci. USA, 77*, 6349–6353.
37. Stoffel, W., Blobel, G. & Walter, P. (1981) *Eur. J. Biochem. 120*, 519–522.

38. Anilionis, A., Wunner, W. H., & Curtis, P. J. (1981) *Nature (Lond.) 294*, 275–278.
39. Min Jou, W., Verhoeyen, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barber, C., Carey, N. & Emtage, S. (1980) *Cell, 19*, 683–696.
40. Goeddel, D. V., Leung, D. W., Dull, T. J., Gross, M., Lawn, R. M., McCandliss, R., Seeburg, P. H., Ullrich, A., Yelverton, E. & Gray, P. W. (1981) *Nature (Lond.) 290*, 20–26.
41. Lawn, R., Adelman, J., Dull, T. J., Gross, M., Goeddel, D. & Ullrich, A. (1981) *Science (Wash. DC) 212*, 1159–1162.
42. Gray, P. W. & Goeddel, D. V. (1982) *Nature (Lond.) 298*, 859–863.
43. Taniguchi, T., Mantei, N., Schwarzstein, M., Nagata, S., Muramatsu, M. & Weissman, C. (1980) *Nature (Lond.) 285*, 547–549.
44. Schechter, I., Wolf, O., Kantor, F., Schechter, B. & Burstein, Y. (1980) *Ann. N.Y. Acad. Sci. 343*, 218–231.
45. Givol, D., Zakut, R., Effron, K., Rechavi, G., Ram, D. & Cohen, J. B. (1981) *Nature (Lond.) 292*, 426–430.
46. Sims, J., Rabbitts, T. H., Estess, P., Slaughter, C., Tucker, P. W. & Capra, J. D. (1982) *Science (Wash. DC) 216*, 309–310.
47. Carne, T. & Schoole, G. (1982) *J. Biol. Chem. 257*, 4133–4140.
48. Gaye, P., Huo, D., Raymond, M.-N., Haze, G. & Mercier, J.-C. (1982) *Biochimie (Paris) 64*, 173–184.
49. Quinto, C., Quiroga, M., Swain, W. F., Nikovits, W. C., Standring, D. N., Pictet, R. L., Valenzuela, P. & Rutter, W. J. (1982) *Proc. Natl Acad. Sci. USA, 79*, 31–35.
50. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N. & Numa, S. (1979) *Nature (Lond.) 278*, 423–427.
51. Kakidani, H., Furutani, Y., Takahashi, H., Noda, M., Morimoto, Y., Hirose, T., Asai, M., Inayama, S., Nakanishi, S. & Numa, S. (1982) *Nature (Lond.) 298*, 245–249.
52. Comb, M., Seeburg, P. H., Adelman, J., Eiden, L. & Herbert, E. (1982) *Nature (Lond.) 295*, 663–666.
53. Joon Yoo, O., Powell, T. & Agarwal, K. L. (1982) *Proc. Natl Acad. Sci. USA, 79*, 1049–1053.
54. Panthier, J.-J., Foote, S., Chambraud, B., Stroseberg, A. D., Corvol, P. & Rougeon, F. (1982) *Nature (Lond.) 298*, 90–92.
55. Rice-Ficht, A. C., Chen, K. K. & Donelson, J. E. (1982) *Nature (Lond.) 297*, 616–618.
56. Maguire, M., Minth, C. D., Funckes, C. L., Deschenes, R., Tavianini, M. A. & Dixon, J. E. (1982) *Proc. Natl Acad. Sci. USA, 79*, 5152–5156.
57. Goodman, R. H., Jacobs, J. W., Chin, W., Lund, P. K., Dee, P. C. & Habener, J. F. (1980) *Proc. Natl Acad. Sci. USA, 77*, 5869–5873.
58. Hobart, P., Crawford, R., Shen, L., Pictet, R. & Rutter, W. J. (1980) *Nature (Lond.) 288*, 137–141.
59. Jacobs, J. W., Goodman, R. H., Chin, W. W., Dee, P. C., Habener, J. F., Bell, N. H. & Potts, J. T. (1981) *Science (Wash. DC) 213*, 457–459.
60. Lund, P. K., Goodman, R. H., Dee, P. C. & Habener, J. F. (1982) *Proc. Natl Acad. Sci. USA, 79*, 345–349.

G. von Heijne, Forskningsgruppen för Teoretisk Biofysik, Institutionen för Teoretisk Fysik, Kungliga Tekniska Högskolan, S-100 44 Stockholm, Sweden

J. Mol. Biol. (1985) 184, 99–105

Signal Sequences
The Limits of Variation

Gunnar von Heijne

Research Group for Theoretical Biophysics
Department of Theoretical Physics, Royal Institute of Technology
S-100 44 Stockholm, Sweden

(Received 29 October 1984, and in revised form 20 February 1985)

Variations in length and composition of the charged N-terminal, central hydrophobic and polar C-terminal regions in a large sample of signal sequences have been mapped, both as a function of the overall length of the sequence, and in an absolute sense, i.e. various "extremes" have been sought. The results show subtle differences between eukaryotic and prokaryotic sequences, but the general impression of signal sequences as being highly variable is reinforced. Criteria for a "minimal" signal sequence are suggested and discussed.

1. Introduction

In the process of protein export a central role is played by the signal sequence: an N-terminal segment that somehow initiates export whereupon it is cleaved from the mature protein. All signal sequences seem to be built along the same general lines, but the fine-structure of the design has only recently become a subject of study (von Heijne, 1983, 1984*a*,*b*; Perlman & Halvorson, 1983). Three structurally dissimilar regions have been recognized so far: a positively charged N-terminal region, a central hydrophobic region and a more polar C-terminal region that seems to define the cleavage site. These regions are present in all signal sequences, but the limits imposed upon them by the export machinery have not been systematically studied; in particular, it has not been ascertained whether they are all equally prone to variations in length and amino acid composition. This is an important question, since one of the outstanding features of the signal sequences taken as a group is their extraordinary variability in terms of overall length and amino acid sequence.

In this paper, eukaryotic and prokaryotic signal sequences are grouped according to their lengths, and the variations with length of the three structural regions (termed the n, h, and c-regions in what follows) are analysed. The analysis shows subtle differences between eukaryotic and prokaryotic sequences, and suggests "minimal" requirements that a fully functional signal sequence must conform to. Available data on non-functional mutant sequences, as well as on export-competent revertants, are discussed in the light of these requirements. Finally, the functional significance of the results are assessed and related to current models of protein export.

2. Methods

The sample under study consists of 118 eukaryotic and 32 prokaryotic signal sequences, all with known cleavage sites. In the prokaryotic sample, no sequences known to be cleaved by the "lipoprotein signal peptidase" (Innis *et al.*, 1984) have been included. Unless otherwise indicated, references can be found in von Heijne (1984*b*).

Prokaryotic sequences: *Escherichia coli* maltose binding protein; phage pBR322 β-lactamase; phage M13 major and minor coat proteins; *E. coli* λ-receptor; *Salmonella typhimurium* histidine binding protein; *S. typhimurium* lysine-arginine-ornithine binding protein; *E. coli* leucine binding protein; *E. coli* leucine-isoleucine-valine binding protein; *E. coli* arabinose binding protein; *E. coli* galactose binding protein; *E. coli* chromosomal β-lactamase; *E. coli* ompA protein; *E. coli* ompF protein; *E. coli* ompC protein; *E. coli* lt enterotoxin A and B-subunits; *Bacillus subtilis* α-amylase; *E. coli* alkaline phosphatase; *E. coli* phoE protein; *Staphylococcus aureus* protein A; *Corynebacterium diphtheriae* toxin tox228; *E. coli* papA; phage Ike gene VIII and gene III proteins; *Vibrio cholerae* toxins ctxA and ctxB (Mekalanos *et al.*, 1983); *E. coli* pilin k88; *E. coli* tolC (Hackett *et al.*, 1983); *E. coli* d-ribose binding protein (Groarke *et al.*, 1983); *Pseudomonas aeruginosa* exotoxin A (Gray *et al.*, 1984); *Pseudomonas* sp. carboxypeptidase G2 (Minton *et al.*, 1984).

Eukaryotic sequences: rat whey phosphoprotein; human serum albumin; rat α_1-acid glycoprotein; mouse thyrotropin α-subunit; atlantic hagfish, anglerfish and human insulins; ovine β and κ-caseins; ovine α and β-lactalbumins; rabbit α-lactalbumin; mouse immunoglobulin H-chain (H-315); hybridoma immunoglobulin H-chain (93σ7); rabbit immunoglobulin H-chain (Pμ3); mouse κ-immunoglobulin L-chains (L-41b, L-315 and

0022-2836/85/130099-07 $03.00/0

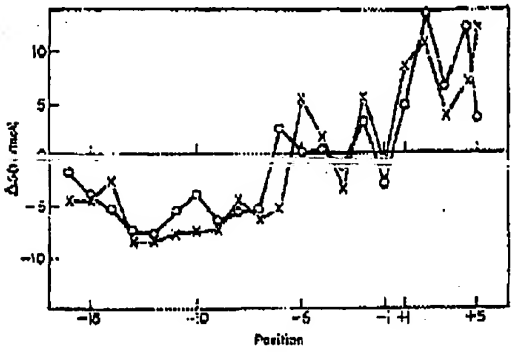


Figure 1. Mean hydrophobicity as a function of the position from the cleavage site (between −1 and +1) for the full eukaryotic (×) and prokaryotic (○) samples. The hydrophobicity scale used is from von Heijne (1981). Negative values are more hydrophobic.

L-321); mouse embryonic V_H-immunoglobulin (PCH 104); immunoglobulin V_H-107; caiman immunoglobulin V_H-gene; conkerel VLDL-II; bee melittin; rat lactin; human placental lactogen; human choriogonadotropin β-subunit; rabbit uteroglobin; rat growth hormone; bovine parathyroid hormone; human leukocyte, immune, γ and fibroblast interferons; rat relaxin; chicken tropoelastin B; chicken ovomucoid; chicken lysozyme; chicken conalbumin; human α_1-antitrypsin; rat prostatic binding proteins C1, C2 and C3; mouse liver amylase; mouse MHC antigen H-2L_d; mouse MHC I-A light chain; mouse MHC I_E-α D chain; mouse MHC E_β chain; human HLA-DR α and DR β_1-chains; human HLA DC-α and DC-β chains; rat carboxypeptidase A; *Torpedo californica* acetylcholine receptor α, β, γ and δ-subunits; human muscle acetylcholine receptor; maize zein protein Z22.1; mouse C3 complement; rat pancreatic RNAase; mouse opiomelanocortin; rat somatostatin; human antithrombin III; rat Thy-1; bovine Arg-Npl1 hormone; yeast invertase; mouse thyrotropin β-subunit; hamster glucagon; pea seed lectin; rat apolipoprotein E; barley α-amylase; human apolipoprotein A I; mouse β-crystallin; rat angiotensinogen; trypsinogen; rat elastase I and II; bovine chymosin; pea legumin; human insulin-like growth factor I; rat seminal vesicle secretion IV protein; rabbit poly-Ig receptor; VS virus glycoprotein G; adenovirus glycoprotein; rabies virus glycoprotein (ERA); human influenza A/Victoria/ and A/Jap/ haemagglutinins; avian influenza A/FPV/ haemagglutinin; *Herpes simplex* virus type-1 glycoprotein D; human parathyroid hormone (Hendy *et al.*, 1981); yeast pho5 (Arima *et al.*, 1983); *Herpes simplex* type 2 glycoprotein D gene (Watson, 1983); rat α-lactalbumin (Qasba & Safaya, 1984); bovine Ot-Npl hormone (Ruppert *et al.*, 1984); human epidermal growth factor receptor (Ullrich *et al.*, 1984); human Christmas factor (Anson *et al.*, 1984); human pancreatic polypeptide (Boel *et al.*, 1984); wheat gliadin (Rafalski *et al.*, 1984); T-cell receptor α-subunit (Saito *et al.*, 1984); human transferrin (Yang *et al.*, 1984); hen ovotransferrin (Williams *et al.*, 1982); *Thaumatococcus daniellii* thaumatin II (Edens *et al.*, [illegible]; human atrial natriuretic factor (Nakayama *et al.*, 1984); human HTLV-I and HTLV-II envelope glycoproteins (Sodroski *et al.*, 1984); human insulin-like growth factor II (Dull *et al.*, 1984); murine epidermal growth factor binding protein (Lundgren *et al.*, 1984); calf acetylcholine receptor γ-subunit (Takai *et al.*, 1984); human α-fibrinogen (Kant *et al.*, 1983); human β and γ-fibrinogen (Chung *et al.*, 1983); human α-haptoglobin (Yang *et al.*, 1983); human retinol binding protein (Colantuoni, 1983); rat chymotrypsin (MacDonald *et al.*, 1982); *Drosophila melanogaster* glue protein 8 (Garfinkel, 1983); VS virus (N. J. Ogden) glycoprotein (Calliono, 1983); *Aplysia* R3-14 neuropeptide (Scheller *et al.*, 1984).

3. Results

(a) *The length distribution is different for eukaryotic and prokaryotic signal sequences*

The number of sequences in the various length-classes is given in Table 1. Except for an extremely short sequence with length $L = 13$, the eukaryotic distribution starts at $L = 15$ (5 sequences) and the prokaryotic at $L = 18$ (3 sequences). Moreover, the main weight of the distributions falls between $L = 18$ and $L = 20$ for the eukaryotes (37%) and between $L = 21$ and $L = 23$ for the prokaryotes (59%). Thus, towards their lower ends the two distributions differ consistently by three residues.

(b) *The mean position of the boundary between the h and c-regions is different in eukaryotes and prokaryotes and does not vary with overall length*

It has been noted that the overall amino acid composition of the c-region is more polar than that of the h-region (von Heijne, 1983). Indeed, in a plot of the mean hydrophobicity of each position in a large enough sample of signal sequences aligned from their cleavage sites, the h/c boundary stands out clearly (Fig. 1). There is an obvious difference between eukaryotes and prokaryotes, however: in eukaryotes, the mean position of the h/c boundary is between residues −6 and −5, whereas in prokaryotes it is between residues −7 and −6. This is true also for the individual length-classes defined

Table 1

Number of signal sequences of given length in the eukaryotic and prokaryotic samples

Length	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	>30
Eukaryotes	1	0	(5	8	4)	13	19	12	9	(3	6)	14	10	(6	1	0	3	2)	2
Prokaryotes	0	0	0	0	0	(3	2	0	7)	(6	6)	(0	4	1	0	0	1)	0	2

The pooled length-classes referred to in the text are enclosed by parentheses when they encompass sequences of more than a single length.

Table 2

Number of residues in the h/c boundary region for a few selected amino acids

Position	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	+1
A. *Eukaryotes*											
Leu	20	10	25	20	27	7	17	5	11	1	5
Val	11	11	9	11	17	2	8	21	2	0	7
Phe	8	6	7	13	13	2	4	2	10	0	4
Ala	11	18	15	12	13	15	15	34	4	53	12
Gly	2	2	8	8	2	15	27	2	4	35	7
Pro	2	0	2	2	6	30	10	1	0	0	1
B. *Prokaryotes*											
Leu	5	11	12	2	0	2	3	1	2	0	0
Val	3	1	5	8	0	0	4	2	0	0	2
Phe	3	2	1	7	1	3	1	0	5	0	0
Ala	11	4	6	5	6	8	5	18	0	28	17
Gly	3	2	5	1	2	4	1	0	1	2	1
Pro	2	1	0	1	4	4	2	0	0	0	0

The h/c boundary (as defined in Fig. 1) is between positions −3 and −5 in the eukaryotes, and between −7 and −6 in the prokaryotes.

in Table 1 (data not shown); thus, the position of the boundary does not vary with overall length. Apart from this difference, the two curves follow each other closely even in the region +1 to +4, indicating that not only the signal sequences proper but also parts of the *mature* sequences are under similar selective pressures (cf. von Heijne, 1984a).

(c) *The h-region is enriched in hydrophobic residues, but has no apparent internal sequence regularities*

The hydrophobic residues Phe, Ile, Leu, Met, Val and Trp are enriched in the h-regions of both eukaryotic and prokaryotic sequences, and drop sharply in frequency at the h/c boundary. Conversely, the charged and polar amino acids (Asp, Glu, Arg, Lys, His, Gly, Pro, Gln, Asn, Ser, Thr and Tyr) are virtually absent in the h-region but dominate the c-region. Ala, which is very abundant in the prokaryotic signal sequences, does not vary appreciably in incidence across the h/c boundary in either sample (Table 2).

It has been claimed that the distribution of amino acids in the h-region is non-random (Inouye & Halegoua, 1980; Perlman & Halvorson, 1983). In the present sample, however, no convincing patterns of fine-structure are apparent in this region; indeed, the number of eukaryotic sequences with a given number of a particular amino acid in the region −13 to −6 closely follows a random expectation, i.e. a binomial distribution (Table 3).

It has also been claimed that some nearest-neighbour pairs of amino acids in the h-region are present in numbers that cannot be explained on the basis of random pairing (Perlman & Halvorson, 1983). This possibility was tested in the present sample by comparing the observed number of pairs of given amino acids in the region −13 to −6 (eukaryotes) or −15 to −7 (prokaryotes) with the numbers obtained for a sample with randomized h-regions (i.e. the amino acids in each individual h-region in the original sample were randomly "scrambled" before the pair-count was performed). As judged by χ^2-analysis (one degree of freedom), no significant deviations ($P < 0.05$) from the expected counts were found, either for nearest-neighbours or for pairs separated by up to three residues, except for Leu-Ile ($i, i+3$)-pairs, which are about twice as numerous as expected (32 *versus* 13) in the eukaryotic sample (data not shown). This indicates that there are no strong sequence constraints in the h-region beyond the observed enrichment in hydrophobic residues.

Table 3

Number of eukaryotic sequences with a given number of Leu, Phe and Gly residues in the region −13 to −6, and the expected numbers calculated on the assumption of a binomial distribution

No. of residues	0	1	2	3	4	5	6	7	8
No. of sequences									
Observed (Leu)	1	8	25	39	34	7	3	1	0
Expected (Leu)	3	11	26	33	27	14	5	1	0
Observed (Phe)	61	39	15	3	0	0	0	0	0
Expected (Phe)	59	43	13	2	0	0	0	0	0
Observed (Gly)	98	13	0	1	0	0	0	0	0
Expected (Gly)	95	21	2	0	0	0	0	0	0

(d) *The n-region accounts for one half of the length variation, but the net N-terminal charge does not vary with length*

The net charge distribution in the n-region differs by one positive charge between eukaryotes and prokaryotes, indicating that the N-terminal amino group in eukaryotes provides one positive charge, whereas the blocked Met_f in prokaryotes does not (von Heijne, 1984*b*). As is clear from Figure 2(a), the net N-terminal charge does not vary appreciably with the overall length in either eukaryotes or prokaryotes, and has a mean value of about +1·7 in both groups. The length of the polar n-region varies strongly with the overall length, however (Fig. 2(b)); the variation is similar in eukaryotes and prokaryotes, and accounts for approximately one half of the total length variation.

(e) *The h-region accounts for one half of the variation in overall length, but there are no regular variations in amino acid composition with length*

Since the length of the c-region is independent of the total length, the remaining half of the length variation stems from the h-region. As is shown in Table 4, there are no regular variations in amino acid frequencies between the different length-classes, and the only suggestive observation so far is

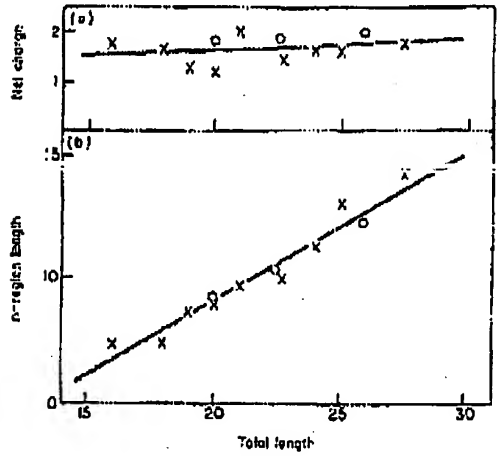


Figure 2. (a) Net charge and (b) length of the n-region as a function of the overall length for the length-classes defined in Table 1 (o, prokaryote; x, eukaryote). The initiator Met is assumed to provide one positive charge in the eukaryotes but not in the prokaryotes (cf. the text). The n/h boundary is defined by either (1) the last charged (Asp, Glu, Arg, Lys) or large polar (Asn, Gln, His) residue, or (2) the last pair of small polar (Ser, Thr, Gly, Pro) residues (whichever yields the longest n-region) on the N-terminal side of the uninterrupted non-polar h-region.

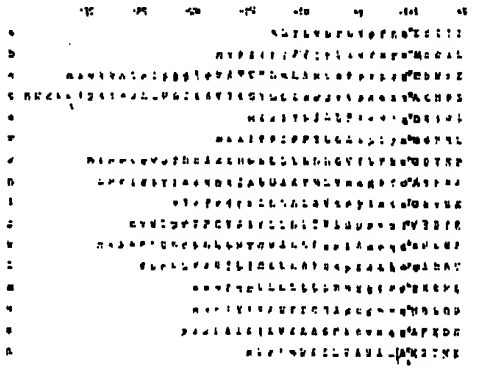


Figure 3. A collection of "extreme" signal sequences (see the text). The h-region is in boldface. Line a, mouse β-crystallin; b, *V. cholerae* toxin (etxA); c, human HLA-DQ β-chain; d, *S. aureus* protein A; e, human influenza A/Jap/ haemagglutinin; f, *E. coli* lt enterotoxin, A-subunit; g, human β fibrinogen; h, phage lke gene VIII-protein; i, chicken α-2 (I) collagen (Tate *et al.*, 1983); j, mouse MHC E_β-chain; k, human pancreatic polypeptide; l, *C. diphtheriae* toxin tox 228; m, human chorionogonadotropin, β-subunit; n, hamster glucagon; o, *E. coli* ompA protein; p, ovine α-S2 casein (Mercier & Gaye, 1980).

that the shortest sequences ($L = 15$ in the eukaryotes, and $L = 18$ in the prokaryotes) have the most hydrophobic h-regions (per residue) both in the eukaryotes and the prokaryotes (the mean hydrophobicity per residue in the region −13 to −6 in the eukaryotic $L = 15$ sequences is −8·9 kJ/mol, the next-largest value is −8·3 kJ/mol (for the $L = 30$ sequences); the value for the region −15 to −7 in the prokaryotic $L = 18$ sequences is −9·6 kJ/mol, the next-largest value is −5·2 kJ/mol (for the $L = 23$ sequences)).

(f) *Examples of "extreme" sequences*

A selection of sequences (extracted from our full collection of some 300 entries) that are "extreme" in one way or another is on display in Figure 3. The first four entries show extremes in overall length, with a more than twofold increase from the shortest ($L = 13$) to the longest ($L = 35$). The variation in the n-region is even more impressive: from one residue (line e) to 17 (line g). The shortest eukaryotic h-region found so far is only seven residues long (line i); the longest is some 16 residues (line k). For the prokaryotes, the corresponding values are nine and 15 (lines h and l). In terms of amino acid composition, there are h-regions that are almost 100% Leu (line m), 0% Leu (line n), rich in Phe (line b), and rich in Ala (line o). The sequence in line d, finally, has an unusually long c-region (10 residues).

(g) *Point mutations, deletions and revertants*

Counting from the cleavage site, almost all export-deficient point mutations described so far

Table 4
Frequencies of a few selected amino acid residues in the h-region of the various length-classes (cf. Table 1)

Length	15–17	18	19	20	21	22–23	24	25	26–30
Leu	0·37	0·39	0·35	0·36	0·43	0·32	0·36	0·29	0·36
Val	0·08	0·13	0·12	0·08	0·10	0·11	0·09	0·16	0·08
Phe	0·08	0·09	0·09	0·13	0·06	0·08	0·07	0·12	0·04
Ala	0·13	0·10	0·09	0·08	0·12	0·10	0·13	0·11	0·09
Ser	0·07	0·04	0·09	0·05	0·05	0·06	0·07	0·04	0·10

The h-region is defined as the region between position −6 and the mean position of the n/h boundary as read from Fig. 2(b).

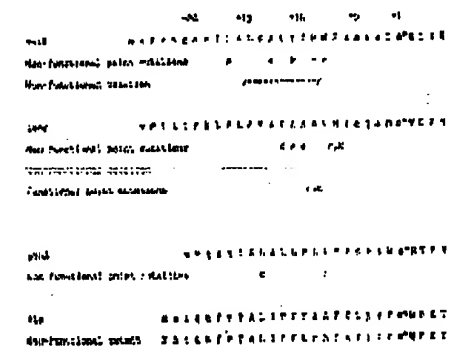


Figure 4. Signal sequence point mutations and deletions (Silhavy *et al.*, 1983). The h-region is in boldface.

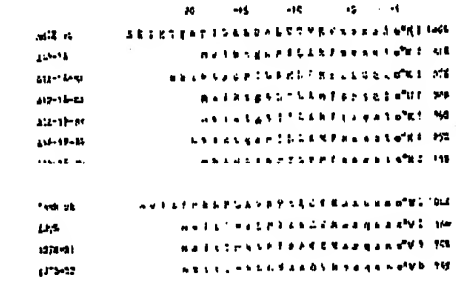


Figure 5. Signal sequence deletions and wholly or partially restored revertants (Bankaitis *et al.*, 1984; Emr & Silhavy, 1983). The h-region is in boldface. The percentage values refer to the amount of correctly exported protein relative to wild-type (wt).

fall in the region −7 to −14 (Fig. 4) except for a $Leu_{-17} \rightarrow Pro$ mutation in malE. Non-functional deletions and partially or totally restored revertants are shown in Figure 5; again, it seems that an intact h-region between residues −7 and −14 (counting from the cleavage site) is required for proper functioning. Exceptions are the functional $Gly_{-9} \rightarrow Arg$ or Asp mutations in lamB, the only indications so far of position-specific effects in the h-region.

4. Discussion

Taken at face value, the results presented here suggest that the n, h and c-regions that together make up the signal sequence are under different selective pressures, although they are by and large similarly selected in eukaryotes and prokaryotes. The c-region does not contribute to the variation in overall length; it extends, with small variations, from residue −1 to −5 (in eukaryotes) or −6 (in prokaryotes), and it follows the "(−3, −1)-rule" for defining the cleavage site (von Heijne, 1983, 1984*a*).

The n-region, on the other hand, is extremely variable both in terms of length and amino acid composition, but its net charge, which is always positive with a mean value of about +1·7, does not vary appreciably with its length.

The h-region, finally, still presents something of an enigma. The statistics on the wild-type sequences and the results from mutation studies agree that residues −7 to −14 (in prokaryotes) or −6 to −13 (in eukaryotes) are the most important ones and constitute what seems to be a "minimal" h-region. Overall hydrophobicity seems to be the one governing principle in this region, and, indeed, substitutions of more hydrophobic for less hydrophobic residues seem to make a big difference in h-regions that are close to the minimal length (Fig. 5): *one* Ser, Gly, Thr or Pro can obviously be tolerated in a "minimal" h-region, but not two. It is uncomfortably true, however, that the lamB Gly_{-9} mutations do not fit this picture without additional assumptions.

Thus, we are left with a picture of the "minimal" signal sequence as one composed of a five (eukaryotes) or six (prokaryotes) residue long c-region; a seven (eukaryotes) or eight (prokaryotes) residue long h-region with at most one Ser, Gly, Thr or Pro among the hydrophobic residues; and a one (eukaryotes) or two (prokaryotes) residue long, positively charged n-region. Thus, it seems that all three regions can be one residue shorter in eukaryotes, making the shortest eukaryotic sequences three residues shorter than the shortest prokaryotic ones (cf. Table 1). If the n, h and c-regions are indeed independent, it should be possible to make a functional 13-residue long eukaryotic signal sequence (cf. Fig. 3, line a) and a 16-residue long prokaryotic one. If the regions are allowed to overlap slightly, even shorter sequences may be possible (see Fig. 3, line p for an example of h/c overlap).

The maximal limits are harder to find. It seems clear that the c-region cannot be much longer than its "consensus" length of five or six residues. Likewise, if the h-region becomes longer than about 20 residues it may in fact anchor the protein permanently to the membrane and turn into an N-terminal *trans*-membrane sequence (cf. von Heijne, 1981; Bos *et al.*, 1984). The limits on the n-region, finally: 18 residues have been attached to the N terminus of a cloned insulin gene, making the n-region 21 residues long, with no effect upon export (Talmadge *et al.*, 1981); on the other hand, a mutant Sindbis virus glycoprotein with its N terminus fused to a 30,000 *M*, cytoplasmic protein is not exported (Wirth *et al.*, 1979). The exact limit remains unknown, however.

What, then, are the functions of these regions? Structurally, they appear to be quite independent, and they do not seem to be co-selected in any important way. The c-region is clearly involved in defining the cleavage site (von Heijne, 1984*a*); the h-region has been suggested as being the target for

the "signal recognition particle" (SRP) which imposes a translational block on cytoplasmic ribosomes synthesizing proteins destined for export (Walter *et al.*, 1981); and the n-region may have something to do with the "docking protein" catalysed release of the SRP-induced block (Hall *et al.*, 1983; Vlasuk *et al.*, 1983). In contrast to the universally conserved "(−3, −1)-design" of the cleavage site (von Heijne, 1983), however, no specific patterns of amino acids have yet been detected in the rest of the signal sequence, and the h as well as the n-region seem far too variable, both in terms of length and amino acid sequence, to allow for specific protein–protein interactions such as those envisioned in the current models of how SRP works. Instead, both regions seem well-suited for binding in a rather unspecific way to the surface (n-region) and to the interior (h-region) of membranes, as suggested repeatedly in the literature (DiRienzo *et al.*, 1978; von Heijne & Blomberg, 1979; Wickner, 1980; Engelman & Steitz, 1981).

The length variations observed in the h-region (~8 to ~20 residues) may indicate that this region spans the 25 to 30 Å thick non-polar interior of the membrane as a structure composed partly of α-helix, partly of extended chain depending on its length: eight residues in a fully extended structure will have a length of some 27 Å, close to the length of a 20-residue long helix. Since the helical conformation should be thermodynamically preferred in a non-polar environment, h-regions of intermediate length would span the membrane as part α, part extended-chain structures (cf. Bedouelle & Hofnung, 1981).

However, this does not explain the SRP-effect. By substituting a polar leucine analogue for Len in a Leu-rich signal sequence one can bypass the SRP-induced translational block and get a cytoplasmic protein; with a Leu-poor signal sequence this does not happen (Walter *et al.*, 1981). It may be, though, that this does not come about as a result of a *direct* interaction between the signal sequence and the SRP; rather, the relation between the two may be indirect and not dependent upon a specific protein–protein interaction. One might speculate that the signal sequence somehow interacts with the ribosome rather than with the SRP, and that the SRP halts translation by interacting with the ribosomal translocation site on ribosomes "sensitized" by the presence of a signal sequence.

This work was supported by a grant from the Swedish Natural Sciences Research Council.

References

Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Gould, K., Huddleston, J. A. & Brownlee, G. G. (1984). *EMBO J.* 3, 1053–1060.

Antoine, M. & Nissing, J. (1984). *Nature (London)*, 310, 798–798.

Arima, K., Oshima, T., Kubota, I., Nakamura, N., Mizunaga, T. & Tohe, A. (1983). *Nucl. Acids Res.* 11, 1657–1672.

Bankaitis, V. A., Rasmusson, B. A. & Bassford, P. J. (1984). *Cell*, 37, 243–252.

Bedouelle, H. & Hofnung, M. (1981) In *Membrane Transport and Neuroreceptors*, pp. 399–403. Alan R. Liss, New York.

[illegible] (1984). *EMBO J.* 3, 909–912.

Bos, T. J., Davis, A. R. & Nayak, D. P. (1984). *Proc. Nat. Acad. Sci., U.S.A.* 81, 2327–2331.

Chung, D. W., Chan, W.-Y. & Davie, E. W. (1983). *Biochemistry*, 22, 3250–3256.

Colantuoni, V. (1983). *Nucl. Acids Res.* 11, 7769–7778.

DiRienzo, J. M., Nakamura, K. & Inouye, M. (1978). *Annu. Rev. Biochem.* 47, 481–532.

Dull, T. J., Gray, A., Hayflick, J. S. & Ullrich, A. (1984). *Nature (London)*, 310, 777–781.

Edens, L., Bum, I., Ledeboer, A. M., Maat, J., Toonen, M. Y., Visser, C. & Verrips, C. T. (1984). *Cell*, 37, 629–633.

Emr, S. D. & Silhavy, T. J. (1983). *Proc. Nat. Acad. Sci., U.S.A.* 80, 4599–4603.

Engelman, D. M. & Steitz, T. A. (1981). *Cell*, 23, 411–422.

Gallione, C. J. (1983). *J. Virol.* 46, 162–169.

Garfinkel, M. D. (1983). *J. Mol. Biol.* 168, 765–789.

Gray, C. L., Smith, D. H., Baldridge, J. S., Harkins, R. N., Vasil, M. L., Chen, E. Y. & Heyneker, H. L. (1984). *Proc. Nat. Acad. Sci., U.S.A.* 81, 2645–2649.

Gronke, J. M., Mahoney, W. C., Hope, J. N., Furlong, C. E., Robb, F. T., Zalkin, H. & Hermodson, M. A. (1983). *J. Biol. Chem.* 258, 12952–12956.

Hackett, J., Misra, R. & Reeves, P. (1983). *FEBS Letters*, 156, 307–309.

Hall, M. N., Gabay, J. & Schwartz, M. (1983). *EMBO J.* 2, 15–19.

Hendy, G. N., Kronenberg, H. M., Potts, J. T. & Rich, A. (1981). *Proc. Nat. Acad. Sci., U.S.A.* 78, 7365–7369.

Innis, M. A., Tokunaga, M., Williams, M. E., Loranger, J. M., Chang, S.-Y., Chang, S. & Wu, H. C. (1984). *Proc. Nat. Acad. Sci., U.S.A.* 81, 3708–3712.

Inouye, M. & Halegoua, S. (1980). *CRC Crit. Rev. Biochem.* 7, 339–371.

Kant, J. A., Lurd, S. T. & Crabtree, G. R. (1983). *Proc. Nat. Acad. Sci., U.S.A.* 80, 3953–3957.

Lundgren, S., Ronne, H., Rask, L. & Peterson, P. A. (1984). *J. Biol. Chem.* 259, 7780–7784.

MacDonald, R. J., Stary, S. J. & Swift, G. H. (1982). *J. Biol. Chem.* 257, 9724–9732.

Mekalanos, J. J., Swartz, D. J., Pearson, G. D. N., Harford, N., Groyne, F. & de Wilde, M. (1983). *Nature (London)*, 306, 551–557.

Mercier, J.-P. & Gayo, P. (1980). *Ann. N.Y. Acad. Sci.* 343, 232–251.

Minton, N. P., Atkinson, T., Bruton, C. J. & Sherwood, R. F. (1984). *Gene*, 31, 31–38.

Nakayama, K., Ohkubo, H., Hirose, T., Inayama, S. & Nakanishi, S. (1984). *Nature (London)*, 310, 699–701.

Perlman, D. & Halvorson, H. O. (1983). *J. Mol. Biol.* 167, 391–409.

Qasba, P. K. & Safaya, S. K. (1984). *Nature (London)*, 308, 377–380.

Rafalski, J. A., Scherer, K., Metzler, M., Peterson, D. M., Hedgcoth, C. & Söll, D. G. (1984). *EMBO J.* 3, 1409–1415.

Rupport, S., Scherer, G. & Schutz, G. (1984). *Nature (London)*, 308, 554–557.

Saito, H., Kranz, D. M., Takagaki, Y., Hayday, A. C., Eisen, H. N. & Tonegawa, S. (1984). *Nature (London)*, **309**, 757–762.
Schuller, R. H., Kaldany, R.-R., Kreiner, T., Mahon, A. C., Nambu, J. R., Schaefer, M. & Taussig, R. (1984). *Science*, **225**, 1308–1308.
Silhavy, T. J., Benson, S. A. & Emr, S. D. (1983). *Microbiol. Rev.* **47**, 313–344.
Sodroski, J., Patarca, R., Perkins, D., Briggs, D., Lee, T.-H., Essex, M., Coligan, J., Wong-Staal, F., Gallo, R. C. & Haseltine, W. A. (1984). *Science*, **225**, 421–424.
Takai, T., Noda, M., Furutani, Y., Takahashi, H., Notake, M., Shimizu, S., Kayano, T., Tanabe, T., Tanaka, K., Hirose, T., Inayama, S. & Numa, S. (1984). *Eur. J. Biochem.* **143**, 109–115.
Talmadge, K., Broeius, J. & Gilbert, W. (1981). *Nature (London)*, **294**, 176–178.
Tate, V. E., Finer, M. H., Boedtker, H. & Doty, P. (1983). *Nucl. Acids Res.* **11**, 61–104.
Ullrich, A., Coussens, L., Hayflick, J. S., Dull, T. J., Gray, A., Tam, A. W., Lee, J., Yarden, Y., Libermann, T. A., Schlessinger, J., Downward, J., Mayes, E. L. V., Whittle, N., Waterfield, M. D. & Seeburg, P. H. (1984). *Nature (London)*, **309**, 418–425.
Vlasuk, G. P., Inouye, S., Ito, H., Itakura, K. & Inouye, M. (1983). *J. Biol. Chem.* **258**, 7141–7148.
von Heijne, G. (1981). *Eur. J. Biochem.* **120**, 275–278.
von Heijne, G. (1983). *Eur. J. Biochem.* **133**, 17–21.
von Heijne, G. (1984*a*). *J. Mol. Biol.* **173**, 243–251.
von Heijne, G. (1984*b*). *EMBO J.* **3**, 2315–2318.
von Heijne, G. & Blomberg, C. (1979). *Eur. J. Biochem.* **97**, 175–181.
Walter, P., Ibrahimi, I. & Blobel, G. (1981). *J. Cell Biol.* **91**, 545–561.
Watson, R. J. (1983). *Gene*, **26**, 307–312.
Wickner, W. (1980). *Science*, **210**, 861–868.
Williams, J., Elleman, T. C., Kingston, I. B., Wilkins, A. G. & Kuhn, K. A. (1982). *Eur. J. Biochem.* **122**, 297–303.
Wirth, D. F., Lodish, H. F. & Robbins, P. W. (1979). *J. Cell Biol.* **81**, 154–162.
Yang, F., Brune, J. L., Baldwin, W. D., Barnett, D. R. & Bowman, B. H. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 5875–5879.
Yang, F., Lum, J. B., McGill, J. R., Moore, C. M., Naylor, S. L., van Bragt, P. H., Baldwin, W. D & Bowman, B. H. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 2752–2756.

Edited by R. Huber

Volume 12 Number 13 1984 Nucleic Acids Research

Compilation of published signal sequences

Marion F.E.Watson

Department of Molecular Genetics, G.D.Searle and Co. Ltd., PO Box 53, Lane End Road, High Wycombe, Bucks, UK

Received 27 April 1984; Revised and Accepted 11 June 1984

INTRODUCTION

The process by which proteins are transported across membranes was presented as a detailed hypothesis, the 'signal hypothesis' in 1975 (1) and subsequently in revised forms (2,3). An essential feature is the presence of a polypeptide of 20-40 amino acids, predominantly hydrophobic, at the amino terminus of secreted proteins. This binds to the endoplasmic reticulum and transport is initiated before translation is complete. Despite some contradictory evidence (e.g. 4-6) the essentials of the signal hypothesis have been widely accepted and have been extended to include proteins which become integrated into membranes (3,7) and bacterial systems (8).

The current hypothesis requires that the signal peptide must be transient. However, a number of secreted and membrane-bound proteins have now been shown to have N-terminal sequences which resemble the transient peptides, but which are not cleaved. It has been shown that these sequences can function as signals for translocation into the rough endoplasmic reticulum (9,10). Therefore, they have been included in this compilation.

Proteins which are transferred across two membrane bilayers into mitochondria and chloroplasts have also now been found to have transient amino terminal regions (3,11,12). A major difference in the transport of these proteins is that it occurs post-translationally and it is clear that their signals are quite different.

The structure, function and processing of signal sequences have been reviewed (7,8,13,14). There has been considerable interest as to which features of signal sequences are essential for their function, notably their secondary structure (15), hydrophobicity (16-19) and the positions of positively charged residues (20,21). The site of cleavage of a signal from its mature protein is of particular interest and a common format for the cleavage site of both prokaryotic and eukaryotic signals has been proposed (22-24). The enzyme responsible for the cleavage step in E. coli and its nucleotide sequence have been identified (25). This protein is itself one of the exceptions to the signal hypothesis in that its signal is not cleaved. The use of cloned DNAs with altered signal sequences (e.g. 26-33) and of cells with mutations in their secretory apparatus (e.g. 34-37) will help answer many of the unsolved questions on the process of secretion.

The following is a compilation of published signal sequences. The proteins are grouped mainly into genera. Phage and plasmid encoded proteins are grouped with their hosts. They are classified as inner membrane, outer membrane, periplasmic or transmembrane proteins where appropriate. The first ten amino acid residues of the following protein sequences are also given. The end of the signal may not always represent an in vivo cleavage site. Where there is clear evidence to the contrary this is indicated at the junction (see abbreviations). Note that each protein complete with its signal peptide should be termed 'pre' protein. In many cases, e.g. the human prefibrinogens, the signal sequence is derived only from DNA sequence data and the actual initiator methionine of several present is unknown. For some proteins the exact end of the signal is unknown. It is inferred by comparison with well characterised junctions using the method of 'processing probabilities' described by von Heijne (24).

The author would welcome any information regarding errors or missing sequences.

of signal sequences
been considerable
uences are essential
structure (15),
positively charge
signal from its
d a common format for
ukaryotic signals has
ple for the cleavage
have been identified
ceptions to the
cleaved. The use of
e.g. 26-39) and of
eratus (e.g. 34-37)
ions on the process

lished signal
y into genera. Phage
ith their hosts. They
mbrane, periplasmic or
The first ten amino
uences are also given.
sent an in vivo
nge to the contrary
reviations). Note
l peptide should be
the human
rived only from DNA
thionine of several
exact end of the
parison with well
of 'processing

4).
tion regarding errors

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
EUKARYOTIC SIGNALS		
BABOON		
α_1-antitrypsin	38	...IILLAGLGGLLPGSLA EDPQGDAAQK
BOVINE		
preparathyroid	39	MMSAKDMVKVMIVMLAICFLARSDG KSVKKRAVSE
growth hormone	40	MHAAGPRTSLLLAFALLCLPWTQVVG AFPAMSLSGL
oxytocin/neurophysin I	41,42	MAGSSLACCLLGLLALTSAFCYIGMCPLGG
vasopressin/ " II	43	MPDATLPACPLSLLAFTSAFCVFGMCPRGG
pro ACTH-β-lipotrophin	44,45	MPRLCSSRSGALLLALLLQASMEVRGWCLESPSAGRDLTTES
glicentin/glucagon	46	MKSLYFVAGLFVMLVQGSWQ HSLQNTEEKS
α-pituitary glycoprotein	47,48	MDFYRKYRAVILTILSLFLGILHS FPDGEFTMQG
protachykinin	49	MKILVAVAVIFFISTQLSAFEEIGAMDDFM
proalbumin	50	MKWVTFISLLLLFSSAYS RGVFRRDTHK
preremin	51	MRGLVVLLAVFALSQG AEITRIPLYK
low MW kininogen 1&II	52	MKLITILFLCSMLLPSLTFQEESQEIDCN
AChR α-subunit	053	MEPRPLLLLLGLCSAGLVLG SEMETRLVAN
CANINE		
trypsinogen I	54	...PLLILAFLAAAVA IPFDDDDKIV
trypsinogen 2 & 3	54,55	MALLFLFLALLLAYVA FPIDDDDKLV PI G
chymotrypsinogen	56	...AFLILVAAFALAAVAFA AAVPAIEPVA
chymotrypsinogen 2	56	MAFLWLLSGFALLGTAFG CGVPAIQPVL
preelastase	58	...FAVAVLAFLLAVAA FAAIPAAPIA
procarboxypeptidase A1	54	...LILVFGALLAAIY GAAFVAAAAA
amylase	54	...FFLLLAVIGFAVA GYAPHAAAAA
preinsulin	57	MALWMRLLPLLALLALWAPAPTRA FVNQHLCGSH
HAMSTER		
preglucagon	58	MKNIYIVAGFFCGAGGSWG HSLQDTEEKS
preinsulin	59	MTLVWMRLLPLLTLVLWEPMPAMA FVNQHLCGSH
HUMAN		
growth hormone	60,61	MATGSRTSLLLAFGLLCLPWLQEGSA FPTIPLSRLF
somatocrinin	62,63	MPLGVFFFVILTLSWE?SHCSPPPDLT
α-gonadotrophin	64	MDYYRKYAAIFLVTLSVFLHVLHS APDVQDCPEC
pre plasminogen activator	65	MDAMKRGLCCVLLLCGAVFVSPSQEIHA?RFRRGARSYN
placental lactogen	66,67,68	MPGSRTSLLLAFALLCLPWLQEAGA VQTVPLSRLF MAA MAP
relaxin	69	MPRLFIFHLLEFCLLLNQFSRAVAA KWKDDVIKLC

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
proinsulin	70	MALWMRLLPLLALLALWGPDPAAA FVNQHLCGSH
insulin like growth fact.	71	(...?)MSSSHLFYLALCLLTFTSSATA GPETLCGAEL
pancreatic peptide	72	MAAAHLCLSLLLLSTCVALLGPLLSAQG APLEDVYPGD
[illegible]	73,74	MDHLCVYVLIFALALAAFSEA?SWKPRSGQPD
α_1 antitrypsin	75,76	MPSSVSWGILLLAGLCCLVPVSLA EDPQGDAAQK
prorenin	77	MDGWRMPRWGLLLLLWGSCIPG?LPTQITTFKR
α-interferon A	78	MALTFALLVALLVLSCKSSCSVG CDLPQTHSLG
N	78	MALTTYLMVALVVLSYKSFSSLG CDLPQTHSLG
C	78	MALSTSLLMAVLVLSYKSICSLG CDLPQTHSLG
H	78	MASPFALLMVLVVLSCKSSCSLG CDLPETHGLD
I	78	...LFLG CDLPQAMSVG
F	78	MALSTSLLMAVLVLSYKSICSLG CDLPQTHSLG
H	78	MALPFSLMMALVVLSCKSSCSLG CNLSQTHSLN
β-interferon	79	MTNKCLLQIALLLCFSTTALS MSYNLLGFLQ
γ-interferon	80	MKYTSYILAFQLCIVLGSLG CYCQDPYVKE
interleukin 2	81	MYRMQLLSCIALSLA?LVTNSAPTSS
Ig heavy chain	482	MEFGLSWLFLVAILKGVQC EVQLLESGGG
Ig κ-chain-IGI	483	MDMRVLAQLLGLLLLCFPGARC DIQMTQSPSS
Ig κ-chain-IG2	483	MDMRVPAQLLGLLLLWLPGAKC DIQMTQSPST
HLA-DR α-chain	484	MIIQKALMLALGALITVMSPCGG EDIVADDHVA
HLA-DR α-chain	485,86	MAISGVPVLGFFIIAVLMSAQESWA IKEEHVIIQA
HLA-DR β-chain	487	MVCLKLPGGSSLAALTVTLMVLSSRLAFA GDTRPRFLEL
HLA-DR β-chains 364	488	MVCLRLPGGSCMAVLTVTLMVLSSPLALA GDTRPRFLEY E
HLA-DR β-chain 5	489	...GGIPGDLRVATVTLMLAILSSSLAEG RDSPEDFVYQ
HLA-CW3	489	MRVMAPRTLILLLSGALALTETWAGSHS?MRYFCIAVSR
HLA-A3	490	MARGDQAVMAPRTLLLLLSGALALTQTWAGSHS?MRYFFTSVSR
apolipoprotein A-I	91,92,93	MKAAVLTLAVLFLTGSQA WFWQDEPP
prothrombin	94	...dLPGCLALAALCSLVHS?GHVFLAPQQA
antithrombin III	95	MYSNVIGYVTSGKRKVVLLSLILIGTWDGVTC HGSPVDICYA
α-fibrinogen	96,97	MFSMRIVCLVLSVVGTAWT ADSGEGDFLA
β-fibrinogen	98,99	MKRMVSWSFHKLKTMVHLLLLLCVFLVKS QGVNDNEEGP
γ-fibrinogen	99	MSWSLHPRNLILYFYALLFLSSTCVA YVATRDNCCI
α-haptoglobin	100,101	MSALGAVIALLLWGQLFA VDSGNDVTDI
prosomatostatin	102	MLSCRLQCALAALSIVLALGCVTGA?PSDPRLRQFL
proenkephalin A	103,104	MARFLYLCTWLLLLGPGLLA?TVRAECSQDC
proenkephalin B	105	MAWDGLVLAACLLMFPSTTA?DCLSRCSLGA

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
pro ACTH-Plipotrophin	106	MPRSCCSRSGALLLALLLQASMEVRG WCLESGQCQD
α-fetoprotein	107	MKWVESIFLIFLLNFTESR TLHRNEVGIA
retinol binding protein	108	MKWVWALLLLAAWAAA ERDCRVSSFR
AChR α-subunit	653	MEPWPLLLLFSLCSAGLVLG SETCTNLVAK
MONKEY		
preinsulin	107	MALWMRLLPLLALLALWGPDPAPA FVNQHLCGSH
MURINE		
Ig heavy chain-315	110,111	MKVLSLLYLLTAIPHIMS DVQLQESGPG
Ig λ_1 chain-104E	112,113.114	MAVISLIGLLALSSGAIS QAVVTQESAL
Ig λ_1 chain-351	113,114	MAVTSLIGLLALCSGASS QAVVTQESAL
Ig λ chain-315	111,113	MAWTSLIGLIALCSGAS QAVVTESALT
Ig κ chain-41A	112,114	MDMRAPAQIFGFLLLLFPGTRC DIQMTQSPSS
Ig κ chain-41B	112	MRAPAQIFGFLLLFPGTRC DIQMTQSPSS
Ig κ chain-321 & 42	114,115	METDTLLLWVLLLWVPGSTG DIVLTQSPAS M
Ig κ chain-Z	116	MSVLTQVLALLLLWLTGARC DIQMTQSPAS
Ig μ-heavy chain	117,118	MKLVLNVVFLITLMGIMC EVKLVESGGG
MLA H-2Kd	119,120	MAPCTLLLLLAAALAPTQTRA GPHSLRYFVT
MLA I-Ek	121	MATIGALVLVFTFIAVLMSSQKSVA IKEEHTIIQA
MLA I-Ed	122	MVVLRRVPCVAAVILLTVLSPPVAL VRDTRPRFLE
epidermal growth factor	123	MPWGRRPTWLLLAFLVFLKISILSVTA?VDTGNCIDFP
mast cell growth factor	124	MVLASSTTSHIMLLLLLMLPMLGLQA?SIRGRDTHRL
pro ACTH-Plipotrophin	125,126	MPRFCVSRSGALLLALLLQTSIDVWS WCLESGQCQD
α_1-interferon	127	MARLCAFLMVLAVMSYWPTCSLG CDLPQTHNLR
α_1-interferon	127	MARLCAFLVMLIVMSYWSICSLG CDLPHTYNLR
β-interferon	128	MNNRWILHAAFLLCFSTTALS INYKQLQLQE
δ-interferon	129	MNATHCILALQLFLMAVSG CYCHGKIVIES
interleukin 3	130	MVLASSTTSIHIMLLLLLMLPMLGLQA?SIRGRDTHRL
α-fetoprotein	131	MKWITFASLILLMVAASK ALMEKEFGIA
complement component 3	132	MGPASGSQLLVLLLLLASSPLALG IPMYSIITPN
properdin	133	MDRRRMPLWALLLWSPCTPSLPTSTT?PESDPLKKHP
submaxillary gland pr.3	134	MKFLALIVLLGVSTILVS?GQDPETNSTE
OVINE		
β-casein	135,136	MKVLILACLVALALA REQFPLNVVS
αS_1-casein	135,136	MKLLILTCLVAVALA RVKHPIKHLG
αS_2-casein	136	MKVLMKACLVAVALA KxxMxxVxxx
αx-casein	135,136	MRKSILLVVTILALTIPFLIA QECMQEQRIC

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
β-lactoglobulin	136	MKGLILALGLALACGVQA IIVTQTMKGL
α-lactalbumin	137	MMSFVSLLLVGILFWATQA EQLTKCEVFQ
ACTH releasing factor	138	MRLPLLVSVGVLLVA?LLPSPPCRAL
transelastin	139	[illegible]
PORCINE		
growth hormone	140	...TSVLLA?ALLCLPWTQEVG AFPAMPLSSL
preproenkephalin B	141	MAWQGLLLAACLLVLPSTMA?DCLSGCEICA
pro ACTH-β-lipotrophin	142	MPRLCGSRSGALLLTLLLQASMGVRGW CLESSQCQDL
relaxin	69	MPRLFSYLLGVVLLSQLPREIPG QSTNDFIKAC
progastrin	143	MQRLCATVLIHVLALAAGSEA?SWKPGFQNSD
α-lactalbumin	144	MMSFVSLLVVGILFPAIQA?KQFTKCELSQ
RABBIT		
IgG VH2Weif2,1S	382	METGLRWLLLVAVLKGVQC QSVKESEGGL
Ig κ-Chain-b5	3165	...TRAPAQLLGLLLLWLPGARCA DVVMTQTPAS
uteroglobin	146	MKLAIFLAVALALLGSPASA GICPRFAHVI
cytochrome P-450	3147	MEFSLLLLIAFLAGLLLLLFRS/KPRAKGKLPP
α-lactalbumin	164	MMPLVPLLLVSIVFPSIQA?TQLTKCELTE
RAT		
trypsin 1	148	MSALLILALVGAAVA FPLEDDDKIV
trypsin II	149	...LVGAAVA FPVDDDDKIV
α_1-antitrypsin	149	MAPssssqLLLLAvLCLAPxxLA DAxxDxxxx
elastase 1	148,150	MLRFLVFALLVLVGMS TQDIPETNAD
elastase II	148,150	MLRTLILSAFVAGALS CGYPTVEVQM
chymotrypsin	148	MAFLWLVSCFALVGATFG Cxxxxxxxx
procarboxypeptidase A	148,151	MKSLLILSLILEAVCG NENPVGHQVL
amylase	148,152	MKFVLLLSLIGFGWA QYDPHTADSR
proinsulin	153,154,155	MALWMRFLPLLALLVLWEPKPAQA FVKQHLCGPH
ribonuclease A	148,156	MGLEKSLFLFSLLVLVLGWVQPSLG GESRESSADK
growth hormone	157	MARDSQTPWLLTFSLLCLLWPQEAGA LPAMPLSSLF
angiotensinogen	158	MIPTGAGLKATIPCILTWVSLTAG DRVYIHPFHL
oxytocin/neurophysin I	159	MAIPSLACCLLGLLALTSA?CYIQNCPLGG
vasopressin/ " II	160	MLAMCLNFTLSACFLSLLALTSA?CYFQNCPRGG
relaxin	161	MSSHLLLGLLGTVLFLSQPCRA RVSEEWMDQV
preproatestatin	162	MLSCRLGCALAALCIVLALGGVTEA?PSDPRLRDFL
prostatein C1	163	MSTVLSLCLLIMLAVCGYEAMA?SQICELVAHE
prostatein C2	163	MRLSLCLLTILVVCGYEAMG?QTLAGVGGAL

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
prostatein C3	164	MKLVPLFLLVTIPICGYA SGHGCSILDE
proalbumin	165,166	MKWVTFLLLLFISGSAFS RGVFRREAHK
prochalecystokin	167	MKCGVCLGVVMAVLAASALAIQPVVPVEAVD
prolactin	168	MKSQVSARKAGTLLLLPMSMLLFGMVHT LPVCSGGDCD
whey phosphoprotein	169	MKCFISLVLGLLALZVAEA RMLQEMVPES
alarmglobulin	170	MKLAITLALVTLALLCSPASA GICFNPAxVI
haptoglobin	171	MxALxAVIxLLLxxQPA ADFSNEVTDI
α_1-acid glycoprotein	172	MALMMVLVVLSLLPLLEA QMPEPAMITL
α_1-interferon	173	MARLCAFLMSLVVVSYWSACCLG CDLPMTHNLR
apolipoprotein A	174	MKAAVLAVALVFLTGSQA AEPxQDDEPQ
apolipoprotein E	175	MKALWALLLVPLLTGCLA EGELEVTPQL
seminal vesicle pr. IV	176	MKSTSLFLGSLLLLLVTGAIG SKTKxxxxxx
procalcitonin	177	MSFLKFSPFLVVSILLLYQACGLQAPVPLRSTLESS
lutropin	178	MERLQGLLLVLLLSPSVVWA SHGPLRPLCR
cytochrome P-450	179,180	MEPSILLLLALLVGFLLLLVRG/xxxxxxxxxx T
cytochrome P-450a	180	MLDTGLLLVVILATLTVMLLLTL/xxxxxxxxxx
CHICKEN		
proinsulin	57,181	MALWIRSLPLLALLVFSGPGTSYA AANQHLCGSH
lysozyme	0182	MRSLLILVLCFLPLAALS XVFGRCELAA
ovomucoid	183	MAMAGVFVLFSFVLCGFLPDAAFG AEVDCSxxFP
conalbumin	185	MKLIICTVLSLGIAAVCFA APPKSEVIRWC
vitellogenin	184	MRGIILALVLTLVGSIQKFDIDPGFN
pro α2(I) collagen	185	MLSFVDTRILLLLALTVTSYLATSQH
pro α1(I) collagen	186	MFSFVxSRLLLLIATAxxLLxxxxx
Apo VLDL-II	186	MQYRALVIAVIILLSTTVPEVCS KSIIDRERRD
Ig ϵ-chain	187	...SGSLVQA ALYQPASVSA
trapoolastin b	188	MKQSAALPLLPGVLLLFSILPASQG GGVPGAIPGG
CAIMAN CROCODYLUS		
Ig(v)heavy chain	4118	MGLGLHLLVLAAALQGAWS QVQLVESGGD
XENOPUS		
neurotensin like peptide	189	MYKGIFLGVLLAVICAMSLAITPSSDADEDN
ANGLER FISH		
proinsulin	185,190	MAALWLQSFSLLVLLVVSWPGSQA VAPAxxLxxx
preglucagon	59,191	MKRIMSLAGILLVLGLIQSSC?RVLMEKADPS
preamelotelatin	102	MKMVSSSRLRGLLVLLLSLTASISGSFA?GMRGSKLRLL

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
CARP		
proinsulin	190	MAVGIGAGALLFLLAVSGVNA HAGAPQMLCG
CATFISH		
preosmatostatin	191	MSSSPIRLALALMCLVEAVGVISFCGRPMVVLNS
CHICKEN		
proinsulin	185,190	MALSPFLAAVIPLVLLLSRAPPSADT RTTGAxxxxx
SALMON		
proinsulin	165	MRLWLQAASLLVLLALSPGVDA AAADMLGSSM
TORPEDO		
AChR α-subunit	6193,194	MILCSYWMVGLVLLLFSCGSLVLG SEMETKLVAN
WINTER FLOUNDER		
antifreeze protein I	195	MDITEANPLMDAKAVPAAAAP DTASDAAAAA
antifreeze protein II	196	MALSLFTVGGLIFLFVTIMGITEASPDP
BEE		
promelittin	197,198	MKFLVNVALVFMVVYISYIYA APEPEPAPED
DROSOPHILA MELANOGASTER		
68C glue protein-6	199	MKLLVVAVIACIMLIGPADPASG GKDGSCVICG
68C glue protein-7	199	MKLIAVTIIACILLIGFSDLALG GASECGPCGP
68C glue protein-3	199	MKLTIATALASILLIGSAMVANC CDCGCPTTTT
yolk protein 1	200	MNPMHVLSLLACLAVAFALAXPMGHMD
BRASICA NAPUS		
napsin	201	MANKLFLVSATLAFTFLLTNAFSIYRIVVEPG
HORDEUM VULGARE		
α-amylase	202	MGKNGSLGGFSLLLLLLAGLASFGMQVLFQGFN
PHASEOLUS VULGARIS		
phaseolin	203	MNGNARVPLLLLGILFLASLSAATAITSLGEEEESG
PISUM SATIVUM		
vicilin	204	MLLAIAFLASVCVSS RSDQEMPFIC
prolectin	205	MASLETEMISFYAIFLSILITYILPFKVWS TETYSPLITK
ZEA MAIZE		
zein protein 19	206,207	MAAXIFCLIMLLGLSASAATA SIFPQCSQAP
zein protein 22.1	208	MATKILALLALLALLVSATNA FIIPQCSLAP
zein protein 22.3	209	MATKILSLIALLALFASATNA SIIPQCSLAP
SACCHAROMYCES		
pro α factor 1	9,209	MRFPSIFTAVLFAASSALA/APVNTTTEDE
pro α factor 2	210	MKFISTFLTFILAAVSVIAS?SDEDIAQVPA

SIGNAL	JUNCTION
AGALLFLLAVSSVNA	HAGAPQHLCG
[illegible]	
LVLLLSRAPPSADT	RTTGIRRAKR
SLLVLLALSPGVDA	AAAGHLCGSH
LVLLLFSCCGLVLG	SEHETRLVAK
PRPDAKAVPAAAAP	DTASDAAAAA
LFTVGGLIFLFVT?	MRITEASPDP
LVFMVVYISYIYA	APEPDPAPEP
QIMLIGFADPASQ	GKDCSCVIGG
SILLIGFSDIALG	GAGECQPCGP
TILLIGSANVANC	GDCGCPTTIT
RVLSLLAGLAVA?	ALAKPEGRMD
ATLAFTFLLTNA?S	IYRTVVPFD
LLLLLLLASLAS?	GHGVLPQREN
QLFLASLSASFA?	TSLRKKEERQ
IAFLASVCVSS	RSDQENPFIF
LTTILFFKVNS	IEYTSFLITK
LLGLSASAATA	SIFPQCSQAP
LALLVNATNA	PIIPQCSLAP
LALFASATNA	SIIPQCSLAP
VLFAASSALA	APVKTTTEDE
ILAAVSVTAS	FSDEQIAQVPA

SOURCE & PROTEIN	REFS.	SIGNAL	JUNCTION
acid phosphatase	#211	MFKSVVYSILAASLAHA	GTIPLGKLAD
invertase	#212	MRLQAFLFLLASFAAKISA	SMKNETSDRP
PLASMODIUM KNOWLESI			
[illegible]	#213	[illegible]	
TRYPANOSOMA BRUCEI			
surface glycoprotein 117	#214	(...?)MLTLSLLYAITPADG	AKEALEYKTV
surface glycoprotein 221	#214	(...?)MPSNQEARLFLAVLVLAQVLPIIVDS	AAEKGFKQAF
VIRAL SIGNALS			
ADENO VIRUS 2			
19K glycoprotein	#215	MGYMILGLIALAAVCSAA	KKVCFKEPAC
HERPES SIMPLEX			
glycoprotein D-1	216,217	MGGTAARLGAVILFVVIVGLHGVRG	KYALADASLK
glycoprotein D-2	218	MGRLTSGVGTAALLVVAVGLRVVCA	KYALADPSLK
AVIAN INFLUENZA			
A/Rostock HA	#219	MNTQILVFALVAVIPTNA	DKICLGHHAV
HUMAN INFLUENZA			
A/Victoria HA	#220	MKTIIALSYIFCLVFA	QDLPGNDNNS
A/Japan HA	#221	MAIIYLILLFTAVRG	DQICIGYHAN
A/WSN HA	#222	MKAKLLVLLYAFVAG	DQICIGYHAN
B/Singapore HA	#223	MKAIIVLLMVVTSNA	DRICTGITSS
B/Lee HA	#224	MKAIIVLLMVVTSNA	DRICTGITSS
A/Udorn HA	#225	MKTDKMDKIIIIGSVSLTIA/TICFLMQIAI	
A/PR8 HA	#226	MKPKQRIITIGSICLVVGLIS/LILQIGNII	
A/WSN HA	#10,227	MKFNQKIITIGSICMVVGLIS/LILQIGNII	
B/Lee HA	#228	MLPSTVQTLTLLLTSGGVLLSLYVSAS/LSTLLTSDVL	
MOUSE MAMMARY TUMOUR VIRUS			
envelope glycoprotein	#229	MNLAPIKEKTAWQHLQALISEAKEVLKTSQTPQNSc LILPLALLSVLGPPPVTQ	EFIVAYLPKP
RABIES VIRUS			
CVS glycoprotein	#230	MVPQVLLFVLLLCFSLCFG	KFPIYTIPDK
ROUS SARCOMA VIRUS			
envelope glycoprotein	#231	MRRALFLQAFTGYPGKTSKKDSKEKPLATSKKDc PEKTPLLPIRVNYILIIGVLVLCEVIGVRA	DVMLLEQPGM
SIMAN ROTAVIRUS-SA11			
glycoprotein NCVP5	#232	MEKLTDLNYTLSVITLMNNTLNI?ILEDPGNAYT	
VESICULAR STOMATITIS VIRUS - HAMSTER			
glycoprotein G	#233	MKCLLYLAFLFIGVNC	KFTIVFPHNQ

SOURCE & PROTEIN	SIZE.	SIGNAL JUNCTION
VESICULAR STOMATITIS VIRUS - HUMAN		
N.J.Ogden glycoprotein	8234	ML3YLIFALAV3PILG KIEIVFPGMY
Ind.SanJuan glycoprotein	8234	MKCLLYLAFLFIGVNC KFTIVFPHNQ
Cocal glycoprotein	8235	MKFLLLFIxLPLxxxx KFSIVFPGNQ
N.J.Concan glycoprotein	8235	MLxPLIFxLxVxPILx KIEIVFPGHx
Ind.Toronto glycoprotein	8235	MKCLLYLxFLFPxVxG KFTIVFPYNQ
YEAST KILLER		
M1 protein	316,337	MTKPTQVLVRSVSILFFITLLHLVVAFLNDVAGPAET
PROKARYOTIC SIGNALS		
BACILLUS AMYLOLIQUIFACIENS		
α-amylase	238	MIQKRKRTVSFRLVLMCTLLFVSLPITKTSA VNGTLMQYFE
preaubtilisin	239	MRGKKVWISLLFALALIFTMAFGSTS?SAQAAGKSNG
BACILLUS CEREUS		
β-lactamase	+240,241	MMILKMKRMLKIGIGVGILGLSITSLTAFe IGESLGVEAKEKTGDV KMKNQATHEE
BACILLUS LICHENIFORMIS		
β-lactamase	+242	MKLVFSTLKLKKAAAVLLFSCVALAGCANNQTNA SQPAEKNEKT
α-amylase	243	MKQQKRLYARLLTLLFALIFLLPHSAAAA ANLNGTLMQY
BACILLUS SUBTILIS		
α-amylase	244,245,246	MFAKRFKTSLLPLFAGFLLLFHLVLAe GPAAASA ETANKSNELT
BACTEROIDES NODOSUS		
pilin	8247	...FTLIELMIVVAIIGILAAFAIPA/YNDYIARSQA
CORYNEBACTERIUM DITHERIAE		
phage prodiphtheria toxin	248,249	MSRKLFASILIGALLGIGAPPSAHA GADDVVDSSK
ESCHERICHIA COLI		
α-amylase	244	MFAKRFKTSLLPLFAGFLLLFHLVLAGPAAAS AETANKSNEL
h.l.t.-B (human)	250	MKKVKCYVLPTALLSSLGAYLFTALLSSLYAHG APQEITELAe
h.l.t -B (porcine)	251	MKKVKCYVLPTALLSSLYAHG APQTITELCS
h.l.t.-A (porcine)	252	MKNITFIFTILLASPLYA KGDRLYRADS
h.s.t. I	253	MKKLMLAIFISVLSFPSFSQS?IELDSSKEK
h.s.t. II	254,255	MKKNIAFLLASMFVFSIATNAYA?STCSNKKDLC
lipoprotein	8256	MKATKLVLGAVILGSTLLAG CSSNAKIDQL
serine receptor	8257	MLKSIKIVTSLLLVLA?VFGIGLQTSQ
λ-receptor	8258,259	MMITLRKLPLAVAVAAGVMSAQAMA VDTNSYANSQ
ompA	8260,261	MKKTAIAIAVALAGFATVAQA APKDNTWYTG

SOURCE & PROTEIN	REFS.	SIGNAL JUNCTION
ompC	0262	MKVKVLSLLVPALLVAGAANA AEVYNKDGNK
omp E(Ic)	0263	MKKSTLALVVMGIVASASVQA AEIYNKDGNK
ompF	0264	MMKRNILAVIVPALLVAGTANA AEIxxxxxxx
toIC	0265	MKKLLPILIRLSLSGFSSLSQA EMLNQVYQQA
pap pili subunit	0266	MIKSVIAGAVAMAVVSFGVNNA AFTIPGSGSK
β-lactamase -chromosomal	+267	MFRITLCALLITASCSTFAPAPQQIKNDIVH
β-lactamase -plasmid	+268,269	MSIQHFRVALIPFFAAFCLPVFA HPETLVKVKD
tonB	+270	MINTSMTLDLPRRFPWFTLLSVGIHGAVVAGLLTYTSVHQVIEL
alkaline phosphatase	+271,272	MKQSTIALALLPLLFTPVTKA ETPEHPVLEN
phosphate binding pr.	+273	MLVNRTTVATVVAATLSMSAFSVFA EASLTGAGAT
maltose binding protein	+274	MKIKTGARILALSALTTMMFSASALA KIEEGKLVIW
leu-specific binding pr.	+275	MKAKARTIIAGMIALAISHTAMA QDIKVAVVGA
leu-ile-val binding pr.	+276	MNIKGKALIAGCIALAFSNMALA xxxxxxxxxx
arabinose binding pr.	+274	MKxTKLVLGAVILTAGLSxGAxA EMLXLGFLxx
lactose permease	-277	MYYLKNTNTWMFGLPFFFYYFIMGA/VFFFFPIWLH
NADH dehydrogenase	-278	MITFLKKIVIVGGGAGGLEMAT/QLGHKLGHXX
phage M13 major coat pr.	-279,280	MKKSLVLKASVAVATLVPMLSFA AEGDDPAKAA
phage M13 minor coat pr.	-280	MXKLLFAIPLVVPFYSHS AETVESCLAK
leader peptidase	-281	MAMMTALILVIATLVTG/IAWCVDKFFF
ERWINIA AMYLOVORA		
lipoprotein	0282	MNRTKLVLGAVILGSTLLAG CSSNAKIDQL
ENTEROBACTER AEROGENES		
ompA	0283	MKKTAIAIAVALAGFATVAQA APKDNTWYAG
HALOBACTERIUM HALOBIUM		
bacteriorhodopsin	M284	MLELLPTAVEGVS QAQITGRPEW
M. MORGANII		
lipoprotein	0285	MGRSKIVLGAVVLASALLAG CSxxxxxxxx
MORAXELLA NONLIQUEFACIENS		
pilin	0247,284	...FTLIELMIVIAIIGILA/AIAIPAYQDY
NEISSERIA GONORRHOEAE		
pilin	0247,287	...FTLIELMIVIAIVGILA/AVALPAYQDY
PSEUDOMONAS SP.		
pilin	0288	...FTLIELMIVVAIIGILA/AIAIPQYQNY
carboxypeptidase G2	289	MRPSINRTAIAAVLATAFVAGT ALAQKHDKVL
SALMONELLA TYPHIMURIUM		
ompA	0290	MKKTAIAIAVALAGFATVAQA APKDNTWYAG

SOURCE & PROTEIN	RKTS.	SIGNAL JUNCTION
His binding protein	-291,292	MKKLALSLSLVLAFSSATAAFA AIPQKIRIGT
lys-arg-orn binding pr.	-291	MKKTVLALSLLIGLGATAASYA AIPQTVRIGT
SERRATIA MARCESCENS		
lipoprotein	0293	MKRTLLVLGAVILGSMSLAG CSSNAKINIL
SHIGELLA DYSENTERIAE		
ompA	0294	MKKTAIAITVALAGFATVAQA APKDNTWYT
STAPHLOCOCCUS AUREUS		
staphylokinase	295	MLKRSLLFLTVLLLLFSFSTSITNEVSASS
β-lactamase	0296	MKKLIFLIVIALVLSACNSNESNA KELNDLEKKY
VIBRIO CHOLERAE		
toxin subunit A	297	MVKIIFVFFIFLSSFSYA NDDKLYRADS
toxin subunit B	297,298	MIKLKFGVFFTVLLSSAYAHG TPQNITDLCA
MITOCHONDRIAL SIGNALS		
NEUROSPORA CRASSA		
ATP synthase	299	MASTAVLASRLASQMAASAKVAPPAVRVADVSKRc TIQTGSPLQTLKNTQMTSIVSAITRQAFQKRA TSSKIAQANV
SACCHAROMYCES		
cytochrome-C peroxidase	300	MITAVMLLPSLGRTAMKASILYLPSAAAAAc AAAAAITAYSQSMKRSSSSPGGGSNMGc VNKWGKAAALAS TTPLVMVASV
EF-Tu (TufM)	301	MSALLPRLLTRTAFKASGKLLRLSSVISRc TFSQTTTSTYAAAFDRSKP
CHLOROPLAST SIGNALS		
CHLAMYDOMONAS REINHARDTII		
RuBPCase	302	MAVIAKSSVSAAVARPASSSVRPMAALKPAVKc S AAVVAAPAEADD MMVWTPVNNK V
PISUM SATIVUM		
RuBPCase	303	...GFYGGLKPMIGTPVKKVMTIDITSIc TSNGGRVKC MQVWPPIGKK
SOYBEAN		
RuBPCase	304	MASSMISSPAVTTVNRAGAGMVAPFTGLKSc MAGFPVTKTNNDITSIASNGGRVQC MQVWPPIGKK
TRITICUM AESTIVUM		
RuBPCase TS294	305	...VAPTQGLKSTAGLPVSRRSNGASLc SSVSNGGRIRC MQVWPIEGIK
RuBPCase WS4.3	305	MAPAVMASSATTVAPFQGLKSTAGLPVSNRc SRGSLGSVSNGGRIRC MQVWPIExxx
RuBPCase WT	306	MAPAVMASSATTVAPFQGLKSTAGLPISCRSc GSTGLSSVSNGGRIRC MQVWPIExxx

AL JUNCTION
JLAPSSATRAFA AIPGKIRIGT
IGLGATAASVA ALPGTVRIGT
AVILGSRSLAG GSSNAKIMM
ALAGPATVAQA APKNMTVYT
LTVLLLLFSTS?SITMEVSASS
LSACNSNSSKA KELNDLEKKV
GTIFLSSFSYA NDSKLYRADS
VLLSSAVAHG TPDNIYDLCA
AVRVADVSKNc
TTRDAFQKRA TSSEIADANV
LYLFSNAAAAc
SSPGGGSNMQc
MVGKAAALAS ITPLVMVASV
LLRLSSVISRc
TPSQTTTSIYAAAAFDRSKP
PRALKPAVEc
VAAPAEADD MQVWTPVRNK
V
KVNTDITSIc
TSNGGRVAC MQVWPPIGKK
VAPFTGLKSc
ASNGGRVQC MQVWPPIGKK
SRRSNGASLc
VSNGGRING MQVWPIEGIK
AGLPVSRNc
VSNGGRING MQVWPIExxx
GLPISGRSc
SNGGRING MQVWPIExxx

ABBREVIATIONS:
... = incomplete sequence data; # = eukaryotic-membrane protein; @ = bacterial outer membrane protein; = = bacterial inner membane protein; + = periplasmic protein; @ = lysosomal protein, M = transmembrane protein; ? = proposed end of signal; / = N-terminus not cleaved; AChR = acetyl choline receptor; ACTH = [illegible]; Ag = antigen; c = continued; HA = haemagglutinin; h.l.t. = heat labile toxin; h.s.t. = heat stable toxin; Ind. = Indiana; MW = molecular weight; NA = neuraminidase; N.J. = New Jersey; orn = ornithine; pr. = protein; RuBPCase = ribulose-1,5-bisphosphate carboxylase (small subunit); x = unknown amino acid; A = ala; C = cys; D = asp; E = glu; F = phe; G = gly; H = his; I = ile; K = lys; L = leu; M = met; N = asn; P = pro; Q = gln; R = arg; S = ser; T = thr; V = val; W = trp; Y = tyr.

DISCUSSION

A total of 277 sequences are presented here. A detailed analysis of their structure is beyond the scope of this report. However, their most striking similarities include their length, their hydrophobicity and the limited number of amino acids which occur at the carboxyl termini of the signals. The general format of a signal seems to include a charged residue within the first five amino acids, followed by a core of at least nine hydrophobic residues which should be sufficient to span a membrane. A helix breaking residue (glycine or proline) or a large polar residue (notably glutamine), frequently occurs four to eight residues before the cleavage site. The pattern of amino acids near the cleavage sites of 78 eukaryotic and viral proteins has been discussed in detail by von Heijne and a scheme suggested for calculating the probability of processing after each residue (24). The most striking feature of cleavage sites is the presence of amino acids with small, uncharged side chains at the carboxyl termini of the signals. If only those where the cleavage site is well characterised are considered, of 135 of the eukaryotic signals, 69% end with alanine or glycine and a further 24% end with serine, cysteine or threonine. If 16 viral signals are included, these percentages become 71% and 22% respectively. For the 40 well characterised prokaryotic signals 88% end with alanine or glycine and a further 10% with serine, cysteine or threonine.

Processing probabilities have been calculated for each sequence in this compilation (data not shown). For 74% of the eukaryotic and viral proteins the observed cleavage site has the highest processing probability. For a further 11% the site with the second highest probability is cleaved. These probabilities are based on data from eukaryotic and viral proteins only. However, for comparison, they were also calculated for the prokaryotic sequences. 39% of the observed sites were in complete agreement with the predictions. A further 15% are cleaved at the second most probable site. The process by which prokaryotic proteins are exported is more complex than for eukaryotic proteins since the correct targeting of proteins to the inner or outer membranes or to the periplasmic space must occur. Information in the mature protein sequences may be responsible for their ultimate localisation (307-310). Experimental evidence for the mechanisms involved is as yet insufficient and in some cases is contradictory (30,311).

Some of the signal sequences are unusually long, notably those of the Bacilli (31-44 residues) and the viral envelope glycoproteins (53 and 64 residues). In the latter the most hydrophobic regions are towards the carboxyl termini of their signals so it may be that these represent 'internal' signals. The signals responsible for targeting proteins to mitochondria and chloroplasts are also long (42-68 residues). They are very different to the other signals presented. This is not surprising since they must be clearly distinguished from 'export' signals and because their transport is initiated only after translation is complete. It is not possible to describe a general structure from the small sample available.

BIBLIOGRAPHY

The references are presented in the abbreviated form first author, year, journal, volume and page numbers.

1. Blobel G. (1975), J. Cell Biol., 67, 835-851.
2. Blobel G. (1979), Symp. Soc. Exp. Biol., 33, 9-39.
3. Blobel G. (1980), P.N.A.S. USA, 77, 1496-1500.
4. Morlon J. (1983), J. Mol Biol., 170, 271-278.
5. Nakamura M. (1983), Mol. Gen. Genet., 191, 1-9.
6. Naithani J. (1983), Cell, 34, 807-814.
7. Wickner W. (1980), Science, 210, 861-868.
8. Michaelis S. (1982), Annu. Rev. Microbiol, 36, 345-365.

9. Bos T.J. (1984), P.N.A.S. USA, 81, 2327-2331.
10. Julius D. (1984), Cell, 36, 309-318.
11. Neupert W. (1981), Trends Biochem. Sci., 6, 1-4.
12. Schatz G. (1983), Cell, 32, 316-318.
13. Kreil G. (1981), Ann. Rev. Biochem., 50, 317-348.
14. Suominen I. (1983), Int. J. Biochem. 15, 591-601.
15. Austen B.M. (1979), FEBS lett., 103, 308-313.
16. von Heijne G. (1982), Eur. J. Biochem., [illegible]
17. von Heijne G. (1982), J. Mol. Biol, 159, 537-541.
18. Emr S.D. (1983), P.N.A.S. USA, 80, 4599-4603.
19. Finkelstein A. (1983), FEBS lett., 161, 176-179.
20. Inoye S. (1982), P.N.A.S. USA, 79, 3438-3441.
21. Vlasuk G.P. (1983), J. Biol. Chem, 258, 7141-7148.
22. Perlman D. (1983), J. Mol. Biol., 167, 391-409.
23. von Heijne G. (1984), J. Mol. Biol., 173, 243-251.
24. von Heijne G. (1983), Eur. J. Biochem.,133, 17-21.
25. Wolfe P.B. (1983), J. Biol. Chem, 258, 12873-12880.
26. Bedouelle H. (1980), Nature, 285, 78-81.
27. Emr S.D. (1980), Nature, 285, 82-85.
28. Hortin G. (1981), Cell, 24, 453-461.
29. Boeke J.D. (1980), J. Mol. Biol., 144, 103-116.
30. Kornland D. (1982), Cell, 30, 903-914.
31. Hall M.N. (1983), EMBO J., 2, 15-19.
32. Michaelis S. (1983), J. Bacteriol., 154, 366-374.
33. Inouye S. (1983), EMBO J., 2, 87-91.
34. Novick P. (1981), Cell, 25, 461-469.
35. Ito K. (1983), Cell, 32, 789-797.
36. Kumamoto C.A. (1983), J. Bacteriol, 154, 253-260.
37. Shiba K. (1984), EMBO J., 3, 631-635.
38. Kurachi K. (1981) P.N.A.S. USA, 78, 6826-6830.
39. Habener J.F. (1978), P.N.A.S. USA, 75, 2616-2620.
40. Noychik R.P. (1982), Nucleic Acids Res., 10, 71-97.
41. Land H. (1983), Nature, 302,342-344.
42. Ruppert S. (1984), Nature, 309, 554-557.
43. Land H. (1982), Nature, 295, 299-303.
44. Nakanishi S. (1981), Eur. J. Biochem. 115, 429-438.
45. Nakanishi S. (1979), Nature, 278, 423-427.
46. Lopez L.C. (1983), P.N.A.S. USA, 80, 5485-5489.
47. Edwin C.R. (1983), Biochemistry, 22, 4856-4860.
48. Goodwin R.G. (1983), N.A.R., 11, 6873-6882.
49. Nawa H. (1983), Nature, 306, 32-36.
50. MacGillivray R.T.A. (1979), Eur.J.Biochem., 98, 477-485.
51. Noir D. (1982), Gene, 19, 127-128.
52. Nawa H. (1983), P.N.A.S. USA, 80, 90-94.
53. Noda M. (1983), Nature, 305, 818-823.
54. Carne T. (1982), J. Biol. Chem., 257, 4133-4140.
55. Devillers-Thiery (1975), P.N.A.S. USA, 72, 5016-5020.
56. Pinsky S.D. (1983), P.N.A.S. USA, 80, 7486-7490.
57. Kwok S.C.M. (1983), J. Biol. Chem., 258, 2357-2363.
58. Bell G.I. (1983), Nature, 302, 716-718.
59. Bell G.I. (1984), Diabetes, 33, 297-300.
60. Martial J.A. (1979), Science, 205, 602-607.
61. Denoto F.M. (1981), N.A.R., 9, 3719-3730.
62. Gubler U. (1983), P.N.A.S. USA, 80, 4311-4314.
63. Mayo K.E. (1983), Nature, 306, 86-88.
64. Piddas J.C. (1979), Nature, 281, 351-356.

65. Pennica D. (1983), Nature, 301, 214-221.
66. Sherwood L.M. (1979), P.N.A.S. USA, 76, 3819-3823.
67. Barrera-Saldana H.A. (1983), J. Biol. Chem., 258, 3787-3793.
68. Seeburg P.H. (1982), DNA, 1, 239-249.
69. Hudson P. (1983), Nature, 301, 628-631.
70. Bell G.I. (1979), Nature, 282, 525-527.
71. Jensen M. (1983), Nature, 306, 609-611.
72. Boel E. (1984), EMBO J., 3, 909-912.
73. Boel E. (1988), P.N.A.S. USA, 80, 2866-2869.
74. Kato K. (1983), N.A.R., 11, 8197-8203.
75. Leicht M. (1982), Nature, 297, 655-659.
76. Bollen A. (1983), DNA, 2, 255-264.
77. Imai T. (1983), P.N.A.S. USA, 80, 7485-7489.
78. Goeddel D.V. (1981), Nature, 290, 20-26.
79. Derynck R. (1980), Nature, 285, 542-546.
80. Gray P.W. (1982), Nature, 295, 503-508.
81. Taniguchi T. (1983), Nature, 302, 305-310.
82. Bernstein K.E. (1982), Nature, 300, 74-76.
83. Bentley D.L. (1980), Nature, 288, 730-733.
84. Chang H.C. (1983), Nature, 305, 813-815.
85. Das H.K. (1983), P.N.A.S. USA, 80, 3543-3547.
86. Schamboeck A. (1983), N.A.R., 11, 8663-8675.
87. Long E.O. (1983), EMBO J., 2, 389-394.
88. Gustafsson K. (1984), Scand. J. Immunol, 19, 91-97.
89. Sodoyer R. (1984), EMBO J., 3, 879-885.
90. Strachan T. (1984), EMBO J., 3, 887-894.
91. Law S.W. (1983), BiocheR. Biophys. Res. Commun., 112, 257-264.
92. Brewer H.B. (1978), Biochem. Biophys. Res. Commun., 88, 623-630.
93. Karathanasis S.K. (1983), P.N.A.S. USA, 80, 6147-6151.
94. Degen S.J.F. (1983), Biochemistry, 22, 2087-2097.
95. Chandra T. (1983), P.N.A.S. USA, 80, 1845-1848.
96. Rixon M.W. (1983), Biochemistry, 22, 3237-3244.
97. Kant J.A. (1983), P.N.A.S. USA, 80, 3953-3957.
98. Chung D.W. (1983), Biochemistry, 22, 3246-3250.
99. Chung D.W. (1983), Biochemistry, 22, 3250-3256.
100. Yang F. (1983), P.N.A.S. USA, 80, 5875-5879.
101. Raugei G. (1983), N.A.R., 11, 5811-5819.
102. Shen L.P. (1982), P.N.A.S. USA, 79, 4575-4579.
103. Comb M. (1982), Nature, 295, 663-666.
104. Noda M. (1982), Nature, 297, 431-434.
105. Horikawa (1983), Nature, 306, 611-614.
106. Takahashi H. (1981), FEBS lett., 135, 97-102.
107. Morinaga T. (1983), P.N.A.S. USA, 80, 4604-4608.
108. Colantuoni V. (1983), N.A.R., 11, 7769-7776.
109. Wetekam W. (1982), Gene, 19, 179-183.
110. Jilka R.L. (1977), P.N.A.S. USA, 74, 5692-5696.
111. Jilka R.L. (1979), J. Biol. Chem., 254, 9270-9276.
112. Burstein Y. (1977), P.N.A.S. USA, 74, 716-720.
113. Tonegawa S. (1978), P.N.A.S. USA, 75, 1485-1489.
114. Burstein Y. (1977), Biochemistry, 17, 2392-2400.
115. Smith G.P. (1978), Biochem. J., 171, 337-347.
116. Nishioka Y. (1980), J. Biol. Chem., 255, 369-374.
117. Early P. (1980), Cell, 19, 981-992.
118. Litman G.W. (1983), Nature, 303, 349-352.

119. Kvist S. (1983), EMBO J., 2, 245-254.
120. Lalanne J.L. (1983), N.A.R., 11, 1567-1577.
121. Hyldig-Nielsen J.J. (1983), N.A.R., 11, 5855-5871.
122. Saito H. (1983), P.N.A.S. USA, 80, 5520-5524.
123. Gray A. (1983), Nature, 303, 722-725.
[redacted]
125. Uhler M. (1983), J. Biol. Chem., 258, 9444-9453.
126. Notake N. (1983), FEBS lett., 156, 67-71.
27. Shaw G.D. (1983), N.A.R., 11, 555-573.
128. Higashi Y. (1983), J. Biol. Chem., 258, 9522-9529.
129. Gray P.W.C. (1983), P.N.A.S. USA, 80, 5842-5846.
130. Fung M.C. (1984), Nature, 307, 233-237.
131. Law S.W. (1981), Nature, 291, 201-205.
132. Wiebauer K. (1982), P.N.A.S. USA, 79, 7077-7078.
133. Panthier J.J. (1982), Nature, 298, 90-92.
134. Windass J.D. (1984), N.A.R., 12, 1361-1376.
135. Gaye P. (1977), Biochem.Biophys.Res.Commun., 79, 903-911.
136. Mercier J.C. (1978), Biochem.Biophys.Res.Commun., 82, 1236-1245.
137. Mercier J.C. (1978), Biochem.Biophys.Res.Commun., 85, 662-670.
138. Furutani Y. (1983), Nature, 301, 537-540.
139. Davidson J.N. (1982), Arch.Biochem.Biophys., 218, 31-37.
140. Atkinson A. (1983), Symp. Genetics & Biotechnology, 198th Meeting of the Genetical Society, London.
141. Kakidani H. (1982), Nature, 298, 245-249.
142. Boileau G. (1983), N.A.R., 11, 8863-8871.
143. Yoo O.J. (1982), P.N.A.S. USA, 79, 1049-1053.
144. Gaye P. (1982), Biochimie, 64, 173-184.
145. Bernstein K.E. (1983), N.A.R., 11, 7205-7214.
146. Atger M. (1979), Biochem. J., 177, 985-988.
147. Haugen D.A. (1977), Biochem.Biophys.Res.Commun., 77, 967-975.
148. MacDonald R.J. (1982), J. Biol. Chem., 257, 9724-9732.
149. Verbanac K.M. (1983), Arch.Biochem.Biophys., 223, 149-157.
150. MacDonald R.J. (1982), Biochemistry, 21, 1453-1463.
151. Quinto C. (1982), P.N.A.S. USA, 79, 31-35.
152. MacDonald R.J. (1980), Nature, 287, 117-122.
153. Villa-Komaroff L. (1978), P.N.A.S. USA, 75, 3727-3731.
154. Talmadge K. (1981), Nature, 294, 176-178.
155. Sorokin A.V. (1982), Gene, 28, 36-46.
156. MacDonald R.J. (1982), J. Biol. Chem., 257, 14582-14585.
157. Seeburg P.H. (1977), Nature, 270, 486-494.
158. Ohkubo H. (1983), P.N.A.S. USA, 80, 2196-2200.
159. Ivell R. (1984), P.N.A.S. USA, 81, 2806-2810.
160. Schmale H. (1983), EMBO J., 2, 763-767.
161. Hudson P. (1981), Nature, 291, 127-131.
162. Funckes C.L. (1983), J. Biol. Chem., 258, 8781-8787.
163. Parkes M. (1982), Nature, 298, 92-94.
164. Viskochil D.H. (1983), J. Biol. Chem., 258, 8861-8866.
165. Strauss A.W. (1977), J. Biol. Chem., 252, 6846-6855.
166. Sargent T.D. (1981), P.N.A.S. USA, 78, 243-246.
167. Deschenes R.J. (1984), P.N.A.S. USA, 81, 726-730.
168. McKean D.J. (1978), Biochemistry, 17, 5215-5219.
169. Dandekar A.M. (1982), P.N.A.S. USA, 79, 3987-3991.
170. Malsky M.L. (1979), J. Biol. Chem., 254, 1588-1585.

171. Chow V. (1983), FEBS lett., 153, 275-279.
172. Ricca G.A. (1981), J. Bi l. Chem., 256, 11199-11202.
173. Dijkema R. (1984), N.A.R., 12, 1227-1242.
174. Gordon J.I. (1982), J. Biol. Chem., 257, 971-978.
175. McLean J.W. (1983), J. Biol. Chem., 258, 8993-9000.
176. Kendala J.C. (1983), N.A.R., 11, 3169-3186.
177. Jacobs J.W. (1981), Science, 213, 457-459.
178. Chin W.W. (1983), P.N.A.S., 80, 4649-4653.
[illegible] S. (1980), P.N.A.S. USA, 77, 965-969.
180. Botelho L.H. (1979), J. Biol. Chem., 254, 5635-5640.
181. Perler F. (1980), Cell, 20, 555-566.
182. Palmiter R.D. (1977), J. Biol. Chem., 252, 6386-6393.
183. Thibodeau S.N. (1978), J. Biol. Chem., 253, 3771-3774.
184. Geiser M. (1983), J. Biol. Chem., 258, 9824-9830.
185. Tate V.E. (1983), N.A.R., 11, 91-104.
186. Chan L. (1980), J. Biol. Chem., 255, 10860-10863.
187. Reynaud G.A. (1983), P.N.A.S. USA, 80, 4099-4103.
188. Karr S.R. (1981), J. Biol. Chem., 256, 5946-5949.
189. Sures I. (1984), P.N.A.S. USA, 81, 380-384.
190. Hahn V. (1983), N.A.R., 11, 4541-4552.
191. Lund P.K. (1982), P.N.A.S. USA, 79, 345-349.
192. Magazin N. (1982), P.N.A.S. USA, 79, 5152-5156.
193. Sumikawa K. (1982), N.A.R., 10, 5809-5822.
194. Noda M. (1982), Nature, 299, 793-797.
195. Lin Y. (1981), P.N.A.S. USA, 78, 2825-2829.
196. Davies P.L. (1982), P.N.A.S. USA, 79, 335-339.
197. Suchanek G. (1978), P.N.A.S. USA, 75, 701-704.
198. Vlasak R. (1983), Eur. J. Biochem., 135, 123-126.
199. Garfinkel M.D. (1983), J. Mol. Biol. 168, 765-789.
200. Hovemann B. (1981), N.A.R., 9, 4721-4734.
201. Crouch M.L. (1983), J. Mol. Appl. Genet., 2, 273-283.
202. Rogers J.C. (1985), J. Biol. Chem., 260, 8169-8174.
203. Slighton J.L. (1983), P.N.A.S. USA, 80, 1897-1901.
204. Lycett G.W. (1983), N.A.R., 11, 2367-2388.
205. Biggins T.J.V. (1983), J. Biol. Chem., 258, 9544-9549.
206. Geraghty D. (1981), N.A.R., 9, 5163-5174.
207. Pedersen K. (1982), Cell, 29, 1015-1026.
208. Marks D.M. (1982), J. Biol. Chem., 257, 9976-9983.
209. Kurjan J. (1982), Cell, 30, 933-943.
210. Singh A. (1983), N.A.R., 11, 4049-4063.
211. Arima K. (1983), N.A.R., 11, 1657-1672.
212. Taussig R. (1983), N.A.R., 11, 1943-1954.
213. Ozaki L.S. (1983), Cell, 34, 815-822.
214. Boothroyd J.C. (1981), N.A.R., 9, 4735-4743.
215. Ahmed C.M.I. (1982), Gene, 20, 339-346.
216. Watson R.J. (1982), Science, 218, 381-384.
217. Eisenberg R.J. (1984), J. Virol., 49, 265-268.
218. Watson R.J. (1983), Gene, 26, 307-312.
219. Porter A.G. (1979), Nature, 282, 471-477.
220. Min Jou W. (1980), Cell, 19, 683-686.
221. Ward C.W. (1979), Brit. Med. Bull., 35, 64-73.
222. Hiti A.L. (1981), Virology, 111, 113-124.
223. Verhoeyen M. (1983), N.A.R., 11, 4703-4712.
224. Krystal M. (1982), P.N.A.S. USA, 79, 4808-4884.
225. Markoff L. (1982), Virology, 119, 288-297.
226. Fields S. (1981), Nature, 290, 213-217.

227. Hiti A.L. (1982), J. Virol., 41, 730-734.
228. Shaw, M.W. (1982), P.N.A.S. USA, 79, 6817-6821.
229. Redmond S.M.S. (1983), EMBO J., 2, 125-131.
230. Yelverton E. (1983), Science, 219, 614-620.
231. Schwartz D.E. (1983), Cell, 32, 853-859.
232. Both G.W. (1983), J. Virol., 48, 335-339.
233. Limsanna V. (1983), J. Biol. Chem., 253, 8667-8670.
234. Gallione C.J. (1983), J. Virol., 46, 162-169.
235. Kotwal G.J. (1983), Virology, 129, 1-11.
236. Skipper N. (1984), EMBO J., 3, 187-111.
237. Bostian K.A. (1984), Cell, 36, 741-751.
238. Palva I. (1981), Gene, 15, 43-51.
239. Wells J.A. (1983), N.A.R., 11, 7911-7925.
240. Glema A. (1983), N.A.R., 11, 4997-5004.
241. Hezes P.S.F. (1983), FEBS lett., 161, 195-200.
242. Neugebauer K. (1981), N.A.R., 9, 2577-2588.
243. Stephens M.A. (1984), J. Bacteriol., 158.
244. Yang H. (1983), N.A.R., 11, 237-349.
245. Yamazaki H. (1983), J. Bacteriol., 156, 327-337.
246. Ohmura K. (1983), Biochem. Biophys. Res. Commun., 112, 678-683.
247. McKern N.M. (1983), FEBS lett., 164, 149-153.
248. Kaczorek M. (1983), Science, 221, 855-858.
249. Ratti G. (1983), N.A.R., 11, 6589-6595.
250. Yamamoto T. (1982), J. Bacteriol., 152, 586-589.
251. Dallas W.S. (1980), Nature, 288, 499-501.
252. Spicer E.K. (1981), P.N.A.S. USA, 78, 50-54.
253. So M. (1980), P.N.A.S. USA, 77, 4011-4015.
254. Lee C.H. (1983), Infect. Immun., 42, 264-268.
255. Picken R.N. (1983), Infect. Immun., 42, 269-275.
256. Inouye S. (1977), P.N.A.S. USA, 74, 1004-1008.
257. Boyd A. (1983), Nature, 301, 623-626.
258. Emr S. (1980), Nature, 285, 82-85.
259. Hedgepath J. (1980), P.N.A.S. USA, 77, 2621-2625.
260. Movva N.R. (1980), J. Biol. Chem., 255, 27-29.
261. Beck E. (1980), N.A.R., 8, 3011-3024.
262. Mizuno T. (1983), FEBS lett., 151, 159-164.
263. Overbeake N. (1983), J. Mol. Biol., 163, 513-532.
264. Michaelis S. (1982), Annu. Rev. Microbiol., 36, 435-465.
265. Hachett J. (1983), N.A.R., 11, 6487-6495.
266. Baga M. (1984), J. Bacteriol, 157, 330-333.
267. Jaurin H. (1981), Nature, 290, 221-225.
268. Ambler R.P. (1978), P.N.A.S. USA, 75, 3732-3736.
269. Sutcliffe J.G. (1978), P.N.A.S. USA, 75, 3737-3741.
270. Postle K. (1983), P.N.A.S. USA, 80, 5235-5239.
271. Kikuchi Y. (1981), N.A.R., 9, 5671-5678.
272. Inoye H. (1982), J. Bacteriol., 149, 434-439.
273. Nagata K. (1984), J. Bacteriol., 157, 909-917.
274. Bedouelle H. (1980), Nature, 285, 78-81.
275. Oxender D.L. (1980), P.N.A.S. USA, 77, 2885-2889.
276. Wilson V.G. (1980), J. Biol. Chem., 255, 6745-6770.
277. Buchel D.E. (1980), Nature, 283, 541-544.
278. Rogers S.L. (1981), Eur. J. Biochem., 116, 165-170.
279. Sugimoto K. (1977), J. Mol. Biol., 111, 487-507.
280. van Wezenbeek P.M.G.F. (1980), Gene, 11, 129-148.
281. Wolfe P.B. (1983), J. Biol. Chem., 258, 12673-12680.

282. Yamagata (1981), J. Biol. Chem., 256, 2194-2198.
283. Br un G. (1983), Eur. J. Biochem., 137, 495-500.
284. Dunn R. (1981), P.N.A.S. USA, 78, 6744-6748.
285. Vlasuk G.P. (1983), J. Biol. Chem., 258, 7141-7148.
286. Froholm L.O. (1977), FEBS lett., 73, 253-256.
287. Hermodson M.A. (1978), Biochemistry, 17, 442.
288. Sustry P.A. (1983), FEBS lett., 151, 442-445.
289. Hinton N. (1983), Symp. Genetics & Biotechnology, 35th Meeting of Genetics Society, London.
290. Preudl R. (1983), Eur. J. Biochem., 134, 497-502.
291. Biggins C.F. (1981), P.N.A.S. USA, 78, 6038-6042.
292. Biggins C.F. (1982), Nature, 298, 723-727.
293. Nakamura K. (1980), P.N.A.S. USA, 77, 1369-1373.
294. Braun C. (1982), N.A.R., 10, 2367-2378.
295. Sako T. (1983), N.A.R., 11, 7679-7693.
296. McLaughlin J.R. (1981), J. Biol. Chem., 256, 11283-11291.
297. Mekalanos J.J. (1983), Nature, 306, 551-557.
298. Lockman H. (1983), J. Biol. Chem., 258, 13722-13726.
299. Viebrock A. (1982), EMBO J., 1, 565-571.
300. Kaput J. (1982), J. Biol. Chem., 257, 15054-15058.
301. Nagata S. (1983), P.N.A.S. USA, 80, 6192-196.
302. Schmidt G.W. (1979), J. Cell. Biol., 83, 615-622.
303. Coruzzi G. (1983), J. Biol. Chem., 258, 1399-1402.
304. Berry-Lowe S.L. (1982), J. Mol. Appl. Genet., 1, 483-498.
305. Smith S.M. (1983), N.A.R., 11, 8719-8734.
306. Broglie R. (1983), Biotechnology, 1, 55-61.
307. Ichihara S. (1982), J. Biol. Chem., 257, 495-500.
308. Benson E.A. (1983), Cell, 32, 1325-1335.
309. Henning U. (1983), Eur. J. Biochem., 136, 233-240.
310. Kadonga J.T. (1984), J. Biol. Chem., 259, 2149-2154.
311. Pollitt S. (1983), J. Bacteriol., 153, 27-32.